Whole-genome sequencing of rare disease patients in a national healthcare system

The NIHR BioResource, on behalf of the 100,000 Genomes Project

Supplementary Information

Enrolment, research ethics and consent

The NIHR BioResource (NBR) has enrolled 9,742 individuals comprising patients with rare diseases and their close relatives as part of a pilot study for the 100,000 Genomes Project. For this study 15 rare disease domains were approved after review by the Sequencing and Informatics Committee of the NBR. Enrolment of participants started in December 2012 and was completed in March 2017. In addition, samples from a second rare diseases pilot study, coordinated by Genomics England Ltd (GEL) are included together with a number of control samples and samples from the UK Biobank cohort [1]. All together study participants were enrolled in one of 18 domains.

The NBR study was coordinated by the University of Cambridge. Participants were recruited mainly at NHS Hospitals in the UK, but also at overseas hospitals (**Supplementary Table 1 (Enrolment by hospital)**, **Extended Data Figure 1a**). All 13,187 participants provided written informed consent, either under the East of England Cambridge South national research ethics committee (REC) reference 13/EE/0325 or under local IRB approval and governance. Obtaining consent for overseas samples was the responsibility of the respective principal investigators at the enrolling hospitals. The NBR retained de-identified versions of the consent forms from overseas participants and a material transfer agreement was applied to regulate the exchange of samples and data between the donor institutions and the University of Cambridge.

The eighteen domains

The participants were enrolled in one 18 domains (**Supplementary Table 1 (Domain metrics)**). The specifics of the domains are described in the following section.

BPD (Bleeding, Thrombotic and Platelet Disorders)**.** Participants for this domain were enrolled at 31 hospitals according to the approved eligibility criteria. Probands had either a platelet disorder, with or without pathological bleeding; a pathological bleeding disorder not explained by platelet function or coagulation factor defects or multiple thrombotic events at a young age. To enrich for inherited causes probands with a positive family history and/or early onset and/or presence of syndromic features were preferentially selected. In participants with informative pedigrees, where possible, affected and unaffected individuals were recruited. Characteristics of a large portion of participants have been reported [2]. Patients with a BPD of known molecular aetiology (e.g. Haemophilia A and B) were excluded from enrolment, with some exceptions.

CNTRL (Process Controls). A set of DNA control samples from consented participants: 27 healthy individuals from the NBR enrolled under the REC approved study BLUEPRINT (REC 12/EE/0040), 5 healthy individuals enrolled under the REC-approved study GENES & PLATELETS Healthy (REC 10/H0304/65) and 18 patients (9 individuals with BMI > 40 and 9 individuals with partial lipodystrophy) collected under the REC-approved study Inherited platelet conditions (BPD) (REC 10/H0304/66). These samples were used for testing laboratory procedures, data transfer and analysis pipelines.

CSVD (Cerebral Small Vessel Disease). Participants for this domain were enrolled at 10 hospitals according to approved eligibility criteria. In short, patients were eligible if they were suspected to have familial CSVD, based on clinical features such as lacunar stroke or cognitive impairment, at an early age (typically < 60 years), and had changes consistent with SVD on magnetic resonance imaging (MRI) of the brain such as lacunar infarcts and white matter hyperintensities. Data on other typical features of monogenic SVD was also included including migraines, encephalopathy and psychiatric disturbance. Clinical information and MRI results for all recruited patients underwent assessment by a Consultant neurologist and databases of previous stroke admissions were screened. Individuals with causal variants in *NOTCH3* were excluded from enrolment and the cohort was further expanded by a retrospective review of patients who tested negative for *NOTCH3* variants.

EDS (Ehlers-Danlos and Ehlers-Danlos-like Syndromes). Participants for this domain were enrolled at 5 hospitals according to approved eligibility criteria. Patients with EDS subtypes of known molecular aetiology were excluded from enrolment. Patients met the Villefranche criteria for hypermobility EDS [3]. Probands had generalised joint hypermobility, skin hyperextensibility, connective tissue fragility and chronic pain. At the time of recruitment patients had previously undergone ophthalmological and cardiac assessments to exclude Marfan syndrome or other non-EDS hereditary disorders of connective tissue. Participants had a family history suggestive for autosomal dominant inheritance. Where possible, affected and unaffected individuals were recruited.

GEL (100,000 Genomes Project–Rare Diseases Pilot). Enrolment for this second pilot study was coordinated by GEL in partnership with the NBR and the main aims were to test processes and capabilities of National Health Service (NHS) hospitals to enrol individuals with rare diseases and their close relatives within the governance framework of the NHS. Patients with a high likelihood or clear clinical evidence of one of 161 rare inherited disorders and their close relatives (probands and both parents, proband and mother, proband only or larger pedigrees) were enrolled at eight NHS hospitals in England. Patients with known causal mutations were not eligible. Detailed clinical characteristics were collected on the participants using the human phenotyping ontology (HPO). All participants are followed over their life course using electronic health data from primary care (general practitioners), secondary care (hospitals) and relevant registries. The GEL data were primarily used to increase the number of WGS results from genetically independent individuals and to have a larger number of ‘controls’ for the BeviMed association analysis (see below).

HCM (Hypertrophic Cardiomyopathy). Participants for this domain comprise individuals with an unequivocal diagnosis of HCM made at one of four UK specialist NHS centres for HCM care. In short HCM is characterised by primary left ventricular hypertrophy and has an estimated population prevalence of approximately 1 in 500. It is the leading cause of arrhythmia and sudden death in athletes and young adults aged under 35 years. All participants were diagnosed below the age of 70, or affected relatives. Prior to inclusion into the cohort, all patients had undergone clinical genetic testing on an HCM gene panel, with no pathogenic mutation identified.

ICP (Intrahepatic Cholestasis of Pregnancy). Severe ICP is defined as gestational pruritus in association with maternal serum bile acids ≥40 µmol/L. It is associated with adverse pregnancy outcomes, including spontaneous preterm labour, fetal asphyxia and intrauterine death. ICP typically presents in the third trimester. Affected women and their relatives are at increased risk of biliary disease in later life, e.g. drug-induced cholestasis, gallstones, and hepatic fibrosis. The ICP cohort comprises women with disease onset before 32 completed weeks of gestation, that affects approximately 1 in 3,000 pregnant women in the UK. Individual cases with severe, early onset ICP were recruited from 14 UK consultant-led antenatal NHS clinics and from three international units in Argentina, Australia and Sweden. Women were excluded from the study if they had other known causes of hepatic dysfunction such as haemolysis, elevated liver enzymes and low platelets (HELLP) syndrome, preeclampsia, acute fatty liver of pregnancy, acute viral hepatitis, confirmed primary biliary cholangitis or any cause of biliary obstruction on ultrasound. No genetic pre-screening for causal variants in known genes was applied before enrolment.

IRD (Inherited Retinal Disorders). IRD describes a phenotypically heterogeneous group of conditions consequent upon dysfunction and/or degeneration of the neural retina or retinal pigment epithelium, resulting in visual impairment. It is the most common cause of severe visual impairment. Most of the individuals were enrolled at the Moorfields Eye Hospital, London and the remainder at four other NHS hospitals. Most individuals had undergone some previous genetic testing using routine diagnostic approaches and the analysis of the genotyping of a fraction of this cohort has been reported previously [4].

LHON (Leber Hereditary Optic Neuropathy). LHON causes subacute sequential bilateral visual loss which is usually irreversible. Participants with a diagnosis of LHON and a positive test for one of three mitochondrial DNA (mtDNA) mutations in Europeans (m.11778G>A, m.3460A>G, m.81440T>C) were enrolled in this cohort. These mutations are found in ~1 in 300 of the UK population, but the prevalence of blindness due to LHON mtDNA mutations is approximately 1 in 40,000. Environmental factors undoubtedly influence the clinical penetrance, but segregation analyses implicate an interaction between a nuclear modifier locus and the mtDNA mutation. The LHON cohort was included to test the hypothesis that the blindness only occurs when both the nuclear and mtDNA variants are present within the same individual.

MPMT (Multiple Primary Malignant Tumours). Participants for this domain were enrolled through the NHS regional clinical genetics services, mostly (> 95%) in the UK. In each kindred there was a clinical suspicion of a cancer predisposition syndrome, but routine genetic assessment/testing had not identified a genetic cause at the time of enrolment. Most (95%) of cases analysed had developed MPMT (defined as ≥2 primaries by age 60 or ≥3 by 70) but a minority had developed a single primary and had a first-degree relative with MPMT. Tumours in the same tissue type and organ were considered separate primaries if, in the case of paired organs, they occurred bilaterally or if the medical record clearly denoted them as separate. The International Agency for Research on Cancer criteria for defining separate primaries were also used [5]. In 95% of enrolled cases only the DNA of a single pedigree member was analysed.

NDD (Neurological and Developmental Disorders). Participants in this group were adults and children with a previously undiagnosed neurological condition. Referrals were from paediatric and adult neurology and clinical genetics specialities. Conditions included were early and severe epilepsy and encephalopathy, dystonia, spasticity, intellectual disability and metabolic disorders. In many, pre-screening of more common genetic causes of disease were performed but not systematically. The majority were singleton cases but some trios of proband, mother and father were included for paediatric conditions.

NPD (Neuropathic Pain Disorders). Participants with extreme neuropathic pain phenotypes (both sensory loss and gain) were recruited at secondary care clinics located in six NHS hospitals. Participants had to be over 18 years of age with a history of life-style altering sensory disorder (either pain or loss of sensation) for greater than three months. Patients with a known underlying genetic cause of chronic pain (e.g. Fabry's disease [ORPHA:324] and *SCN9A* erythromelalgia [ORPHA:1956]), intellectual disability and/or autistic spectrum disorder sufficient that they could not give either consent or partake in additional pain phenotyping were excluded. The Neuropathic Pain Special Interest Group of the International Association for the Study of Pain grading for neuropathic pain was used for the assessment of all participants [6]. Further details are available in **Appendix 1: NPD** and **Supplementary Table 1 (NPD Criteria – Diagnostic Criteria, NPD Criteria – Outcome Measures)**.

PAH (Pulmonary Arterial Hypertension). In PAH, adverse remodelling of the pulmonary vasculature causes narrowing and obliteration of the capillary arteries in the lung, resulting in elevated resting mean pulmonary artery pressure and right heart dysfunction. A mean pulmonary artery pressure of 25 mm Hg or above, with a pulmonary capillary wedge pressure of less than 15 mm Hg is indicative of PAH. The diagnosis of idiopathic PAH was based on the exclusion of other associated forms of PAH. Participants with idiopathic PAH were recruited from 10 NHS UK National Pulmonary Hypertension hospitals and four international hospitals. All enrolled patients had a clinical diagnosis of idiopathic PAH, heritable PAH, drug-associated PAH, or pulmonary veno-occlusive disease/pulmonary capillary haemangiomatosis (PVOD/PCH) established by their expert centre. The findings in this cohort have been recently reported [7, 8, 9]. No systematic genetic pre-screening for causal variants in previously established genes was applied before enrolment.

PID (Primary Immune Disorders). Participants were recruited by specialists in clinical immunology (either trained in paediatrics or internal medicine) from 21 NHS hospitals in the UK and a small number of hospitals from The Netherlands. In short, the eligibility criteria included the following: clinical diagnosis of common variable immunodeficiency disorder (CVID) according to internationally established criteria [10], extreme autoimmunity, or recurrent (and/or unusual) severe infections suggestive of defective innate or cell-mediated immunity. Patients with known secondary immunodeficiency (e.g. due to cancer or HIV infection) were excluded. Genetic screening for known causes of PID prior to enrolment was encouraged but not applied systematically. Within a broad range of phenotypes, CVID is the most common disease category, comprising 46% of the cohort. Participants for this domain consisted predominantly of singleton cases, but the DNA samples from additional affected and/or unaffected pedigree members of some of the patients were also sequenced. A manuscript describing the findings in this cohort is under review (Thaventhiran et al, under review).

PMG (Primary Membranoproliferative Glomerulonephritis). PMG refers to kidney disease in which a biopsy shows increased glomerular mesangial matrix and cellularity with thickening of the capillary walls and there is an absence of an underlying infectious, neoplastic or autoimmune disorder. Participants in this domain were enrolled from all 10 NHS paediatric renal units in the UK (64 patients) and 18 NHS adult renal centres (120 participants, of whom 21 had paediatric onset of disease). IC-PMG refers to PMG where there is deposition of immunoglobulins and complement C3 in the glomeruli, and 'C3 glomerulopathy' (C3G) is where there is C3 without significant immunoglobulins deposited. The C3G category is further subdivided by electron microscopic appearance into C3 Glomerulonephritis (C3GN) and Dense Deposit Disease (DDD). PMG has an estimated incidence of 3-5 in 1 million and has a poor renal prognosis [11]. Where available, kidney biopsies were reviewed centrally to confirm classification into IC-PMG, C3GN or DDD. No genetic pre-screening for causal variants in known genes was applied before enrolment.

SMD (Stem cell and Myeloid Disorders). Participants for this domain were enrolled if presenting with an inherited bone marrow failure of unknown molecular aetiology presenting in childhood, or an inherited cytopenia (including erythroid lineage) where acquired causes were excluded and without a high index of suspicion for autoimmune aetiology. In addition, patients presenting with a myeloproliferative phenotype and a positive family history of a related haematological disorder were enrolled. Recruitment into the cohort was done by paediatric and adult haematologists from NIHR/National Cancer Research Network Primary Treatment Centres. In addition, cases and their close relatives were enrolled at centres in Canada, Egypt, Norway, Sri Lanka, South Africa, Sweden, Turkey and the USA. The DNA of most probands had undergone previous genetic analysis using targeted sequencing with phenotype-specific gene panel tests and no pathogenic variants were identified.

SRNS (Steroid Resistant Nephrotic Syndrome). The incidence of SRNS is estimated at 2-4 in 100,000 people. There are two major clinical subsets: primary - defined as no response to high dose steroid at four weeks in children or four months in adults, and secondary SRNS where initial steroid sensitivity is lost and treatment resistance evolves as a secondary event either rapidly or over time. Participants for this domain were enrolled through the UK NephroS study, at tertiary NHS paediatric nephrology centres and adult renal units across the UK. Patients with either sporadic or familial SRNS were eligible. Cases with SRNS secondary to systemic disease or obesity were excluded from enrolment. The majority of patients had a histological renal biopsy diagnosis that allowed a second tier of stratification based on histology as well as drug response. The majority (70%) of the cases had undergone previous genetic testing by whole exome sequencing and tested negative for causal variants in diagnostic-grade SRNS genes.

UKB (UK Biobank – Extreme Red Cell Trait Score). The UK Biobank is a biomedical cohort of approximately half a million participants recruited in the UK between 2006 and 2010 [1]. The participants, 54% of whom are women, were aged between 37 and 73 years at their date of recruitment. Each participant underwent a baseline assessment at one of 21 centres distributed across Great Britain, during which EDTA anticoagulated blood was collected for full blood count (FBC) analysis [12]. A subset of UK Biobank participants likely to carry rare alleles modulating erythropoiesis or red cell survival/clearance mechanisms were selected for WGS by considering a univariate score derived from the FBC-measured MCV and RBC# values. 383 and 381 participants from the extreme left and extreme right tails of the score distribution passed quality control after being successfully sequenced. Further details of the approach used for selecting samples are given in **Appendix 2: UKB**.

Overview of sample collection

Genetically-determined sex chromosome status, ethnicity and relatedness for all samples (detailed below) and age of participant at sample collection date were available for the majority of samples. These metrics highlight some of the differences in disease presentation and recruitment approach between the domains (**Extended Data Figure 1b, Supplementary Table 1 (Domain metrics)**). The ICP domain comprised only female participants because the phenotype is related to pregnancy. All domains had a high rate of singletons but for GEL and SMD, for which a large proportion of recruits were part of parent-child trios or mother-child duos. The overall ethnic breakdown for the collections closely matched that of the 2011 UK census [13], suggesting equality of access (**Supplementary Table 1 (Domain metrics)**). In some cases age of diagnosis was available and was similar to age at enrolment, while in others there was a lag of several years between the two. Age at sampling was significantly older than age of presentation of symptoms for domains with a paediatric onset (IRD, NDD, SMD), of childbearing age for ICP, later in life for late-onset disorders (CSVD, HCM, MPMT, PAH) and over the age of 50 for UKB due to its enrolment criteria (**Extended Data Figure 1b**). The sequence data generated on the 13,037 DNA samples (**Supplementary Table 1 (Domain metrics)**) was aggregated and analysed for this manuscript.

Clinical and laboratory phenotype data

Staff at hospitals responsible for enrolment were provided with the eligibility criteria for their respective domains as described above in the domain descriptions. There were different levels of pre-screening amongst the 15 domains to exclude previously known disorders. PAH had virtually no genetic pre-screening. IRD, NDD, BPD, CSVD, EDS, ICP, NPD, PID, PMG and SMD had pre-screening based on clinical presentation and the results of non-DNA based laboratory and imaging tests. Many of IRD, PID, BPD and NDD had received some baseline genetics testing where no pathogenic mutation had been detected. HCM and SRNS had received extensive genetic screening for the relevant diagnostic-grade genes. The clinical and laboratory phenotype data were captured through case report forms (CRF) by paper questionnaires or by online CRF data capture applications and deposited in the NBR study database managed by the University of Cambridge. Online data capture allowed for the entry of Human Phenotype Ontology (HPO) terms [14, 2, 15] by staff at the enrolment hospital and data from paper questionnaires were transformed into HPO terms by the NBR study coordination office. Free text entries were transformed into HPO terms where feasible. An overview of the HPO data obtained for the 15 domains is depicted in **Extended Data Figures 1c and 1d**.

DNA sequencing

Samples were received as either DNA extracted from whole blood or as whole blood EDTA samples that were extracted at a central DNA extraction and QC laboratory in Cambridge. Samples were tested for adequate concentration (Picogreen), DNA degradation (gel electrophoresis) and purity (OD 260/280 quality control (Trinean)) before selection for WGS. DNA samples were prepared at a minimum concentration of 30 ng/µl in 110 µl, visually inspected for degradation and had to have an OD 260/280 between 1.75 and 2.04. They were then prepared in batches of 96 and shipped on dry ice to the sequencing provider (Illumina Inc, Great Chesterford, UK).

Further sample QC was performed by Illumina to ensure that the concentration of the DNA was > 30 ng/µl and that every sample generated high quality genotyping results (Illumina Infinium Human Core Exome microarray). Samples with a repeated array genotyping call rate < 0.99, high levels of cross-contamination, mismatches with the declared gender that could not be resolved by further investigation, or for which consent had been withdrawn, were excluded from WGS (n = 59). The genotyping data were also used for positive sample identification and sample identity was verified before data delivery. In short 0.5 µg of the DNA sample was fragmented using Covaris LE220 (Covaris Inc., Woburn, MA, USA) to obtain an average size of 450 base pair (bp) DNA fragments. DNA samples were processed using the Illumina TruSeq DNA PCR-Free Sample Preparation kit (Illumina Inc., San Diego, CA, USA) on the Hamilton Microlab Star (Hamilton Robotics, Inc, Reno, NV, USA). The final libraries were checked using the Roche LightCycler 480 II (Roche Diagnostics Corporation, Indianapolis, IN, USA) with KAPA Library Quantification Kit (Kapa Biosystems, Inc, Wilmington, MA, USA) for concentration.

From February 2014 to June 2017 three read lengths were used: 100bp (377 samples), 125bp (3,154 samples) and 150bp (9,656 samples). Samples sequenced with 100bp and 125bp reads utilised three and two lanes of an Illumina HiSeq 2500 instrument, respectively. Samples sequenced with 150bp reads utilised a single lane of a HiSeq X instrument. At least 95% of the autosomal genome had to be covered at 15X and a maximum of 5% of insert sizes had to be less than twice the read length. Following sample and data QC at Illumina, 13,187 sets of WGS data files were received at the University of Cambridge High Performance Computing Service (HPC) for further QC.

Analysis of genotyping data

Overview of data processing pipeline

The WGS data for the 13,187 samples returned by Illumina underwent a series of processing steps (**Extended Data Figure 2**), described in detail in the following sections. Briefly, the samples were sex karyotyped and pairwise kinship coefficients were computed. This information was used to check for repeat sample submissions and sample swaps. Additionally, four further QC checks were applied to ensure the single nucleotide variants (SNVs) and short (<= 50bp) insertions/deletions (indels) were of a high standard. Overall, the sequence data from 150 samples (1.1%) were removed, leaving a dataset of 13,037 samples for downstream analysis. The 13,037 individuals were assigned one of the following ethnicities: European, African, South Asian, East Asian or Other. Pairwise relatedness adjusted for population stratification was then computed and used to generate networks of closely related individuals and to define a maximal set of 10,259 unrelated individuals.

The variants in the 13,037 individuals were left-aligned and normalised with bcftools, loaded into our HBase database and filtered on their overall pass rate (OPR), defined below. The sex karyotypes, the ethnicities and the relatedness estimates were used, along with enrolment information, to annotate the samples and variants. Samples were annotated with: affected/unaffected status, membership of the set of probands, membership of the maximal unrelated set, ethnicity and sex karyotype. Variants were annotated with CellBase consequence predictions, HGMD information where available and population-specific allele frequencies.

Alignment and SNV/indel calling

Reads were aligned with the Illumina Isaac aligner version SAAC00776.15.01.27 [16] to reference GRCh37 [17]. This produced BAM files [18] of mean size approximately 95GB for 100 bp samples, 60GB for 125 bp samples and 65GB for 150bp samples. SNVs and indels were called on all BAM files using the Illumina Starling software version 2.1.4.2 [19]. The genotype calls were produced in single-sample mode as both regular VCF and genome VCF (gVCF) files [20]; this second file format also provides information on coverage and alignment quality in homozygous reference regions as well as variant positions. This means that the coverage, alignment quality and other factors allow a PASS filter to be applied to all positions in the genome for all samples.

Sample identity checks

A set of 8,872 common single nucleotide polymorphisms (SNPs) was chosen at the start of the project with which to compute an approximate pairwise kinship between samples as the data arrived at the HPC as a means of identifying duplicate samples. The SNPs were a random subset of those typed by Roche on their targeted microarray platforms for the purpose of inferring kinship. These platforms are used in-house for clinical genomics applications. The kinship coefficients were computed using PLINK [21] after each data delivery. More precise relatedness estimation was performed at a later stage (see below). Any coefficient (PI-HAT) > 0.99 triggered an investigation to determine whether there had been accidental duplicate enrolment, laboratory duplication or enrolment of more than one monozygotic twin. For all pairs resolved to be the same individual, the sample with the highest WGS data quality was retained and the partner excluded from the resource (124 duplicates and 1 triplicate) and the 22 pairs of monozygotic twins were retained in the resource. Eight unconfirmed duplicates were excluded. Most of the duplicate submissions occurred at recruitment due to patients being seen by multiple clinicians. There was one duplicate observed between domains and this patient was confirmed as having a phenotype consistent with the rare disease phenotypes of both domains.

Participant-declared sex and genetically-determined genders were compared for concordance following each WGS data delivery. For each sample, we counted the number of heterozygous, homozygous and hemizygous QC-passing SNVs in the non-pseudoautosomal region of the X chromosome. We defined the H-ratio as the number of heterozygous variants divided by the number of homozygous and hemizygous variants. We made initial gender calls based on the H-ratio as follows: individuals with H-ratio < 0.02 were declared putatively male and the others were declared putatively female (**Extended Data Figure 3e**). Comprehensive sex karyotyping was performed at a later stage (see below). Based on the initial gender calls, two samples with discordant genders were excluded from further analysis.

Variant quality checks

We performed the following variant quality checks:

- We computed the proportion of positions in the non-N fraction of the autosomes that passed quality filters in the gVCFs. This proportion is referred to as the autosomal callability. Samples for which fewer than 95% of bases in the reference genome passed were removed (**Extended Data Figure 3a**).
- We identified a set of common SNPs by downloading and merging the genome and exome data from gnomAD [22] and computing the minor allele frequencies (MAFs) in each population in our study (except for South Asian, which is absent from the gnomAD genome set). The set of SNVs that had a MAF > 5% overall and in each population was recorded. We then determined the proportion of positions in each gVCF that did not pass QC filters as a percentage of the overall set. Samples for which this proportion exceeded 0.55% were excluded (**Extended Data Figure 3b**).
- Transition-transversion ratios (Ts/Tv) are commonly used to determine SNV calling accuracy [23]. We calculated Ts/Tv ratios for all samples and observed that the distribution of these ratios varied depending on the read length batch; samples with a Ts/Tv ratio outside the corresponding batch-specific interquartile range were excluded (**Extended Data Figure 3c**).
- We estimated the degree to which a DNA sample was contaminated by any other DNA sample using verifyBamID [24]. Samples with an estimate of contamination (FREEMIX) exceeding 3% were excluded (**Extended Data Figure 3d**).

In total, 14 samples were excluded due to these four variant quality controls.

Coverage assessment

The mean (across samples) of the mean autosomal coverage by non-duplicated reads was 41.4, 37.9 and 35.3, and the mean 10th percentile of coverage was 31.0, 25.7 and 26.2 for the 100bp, 125bp and 150bp read length batches, respectively (**Figure 1b**). The mean depths were greater than 30X in all samples and 90% of the reference genome was covered at least to 19X in all samples. In addition, all samples were covered to at least 15X in at least 95% of the reference autosomes. The average read duplication rate was 1.5% for read pair inserts in the HiSeq 2500 (100bp and 125bp read lengths), however this value was 18% for the HiSeq X-generated 150bp data, which meant that extra reads were required to make up a similar level of coverage.

Computation of sex karyotypes

The sex karyotypes following delivery of the complete dataset were determined as follows using the BAM and VCF files. For each sample and chromosome, we normalised the number of aligned reads by dividing them by the number of bases which are non-N in the reference genome. The X/Auto and Y/Auto ratios were defined as the normalised read counts on X and Y divided by the median of the normalised read counts on the autosomes (Auto). The median was used because it is robust to large copy number alterations. For the putative males and putative females (defined based on the H-ratio, as described above), the mean and standard deviation of X/Auto and Y/Auto were computed (**Extended Data Figure 3f**). Let us define these values as mean_z(X/Auto), sd_z(X/Auto), mean_z(Y/Auto) and sd_z(Y/Auto), where z = {m,f} denotes whether the summaries were obtained from the putative males or the putative females. We defined the following four gates to classify individuals according to their specific X/Auto and Y/Auto values, denoted x and y, respectively:

- XY gate (shown in blue in **Extended Data Figure 3f**):
    - $\mathrm{mean}_m(X/Auto) - 10\ \mathrm{sd}_m(X/Auto) < x < \mathrm{mean}_m(X/Auto) + 10\ \mathrm{sd}_m(X/Auto)$
    - $\mathrm{mean}_m(Y/Auto) - 10\ \mathrm{sd}_m(Y/Auto) < y < \mathrm{mean}_m(Y/Auto) + 10\ \mathrm{sd}_m(Y/Auto)$
- XYY gate (shown in green in **Extended Data Figure 3f**):
    - $\mathrm{mean}_m(X/Auto) - 10\ \mathrm{sd}_m(X/Auto) < x < \mathrm{mean}_m(X/Auto) + 10\ \mathrm{sd}_m(X/Auto)$
    - $y > \mathrm{mean}_m(Y/Auto) + 10\ \mathrm{sd}_m(Y/Auto)$
- XX gate (shown in red in **Extended Data Figure 3f**):
    - $\mathrm{mean}_f(X/Auto) - 10\ \mathrm{sd}_f(X/Auto) < x < \mathrm{mean}_f(X/Auto) + 10\ \mathrm{sd}_f(X/Auto)$
    - $\mathrm{mean}_f(Y/Auto) - 10\ \mathrm{sd}_f(Y/Auto) < y < \mathrm{mean}_f(Y/Auto) + 10\ \mathrm{sd}_f(Y/Auto)$
- XXY gate (shown in purple in **Extended Data Figure 3f**):
    - $\mathrm{mean}_f(X/Auto) - 5\ \mathrm{sd}_f(X/Auto) < x < \mathrm{mean}_f(X/Auto) + 5\ \mathrm{sd}_f(X/Auto)$
    - $\mathrm{mean}_m(Y/Auto) - 5\ \mathrm{sd}_m(Y/Auto) < y < \mathrm{mean}_m(Y/Auto) + 5\ \mathrm{sd}_m(Y/Auto)$
- XXX gate (shown in orange in **Extended Data Figure 3f**):
    - $1.5\ \mathrm{mean}_f(X/Auto) - 5\ \mathrm{sd}_f(X/Auto) < x < 1.5\ \mathrm{mean}_f(X/Auto) + 5\ \mathrm{sd}_f(X/Auto)$
    - $\mathrm{mean}_f(Y/Auto) - 5\ \mathrm{sd}_f(Y/Auto) < y < \mathrm{mean}_f(Y/Auto) + 5\ \mathrm{sd}_f(Y/Auto)$

One sample had an H-ratio of 0.01 and was thus declared initially as a male. However, the sample was located within the XX gate. Coverage information indicated that there were two X chromosomes, but that they were almost identical in sequence to one another. Further investigation showed that this was due to consanguinity. This and other anomalies are labelled in **Extended Data Figure 3g**. Where possible, within the ethics of the study, non-standard sex karyotypes were confirmed with the referring clinician (n = 5). The observed frequencies of non-standard sex karyotypes, which were present in 13 of the 13,037 individuals (5 XXY, 4 XYY, 2 XXX, 2 XO), were comparable with those reported in a previous study [25]. The variants for the samples with non-standard sex karyotypes (except for XXX) were recalled to take into account the correct ploidy in the sex chromosomes (**Extended Data Figure 2**).

Ancestry and relatedness estimation

We estimated the degree of relatedness between individuals and categorised each individual's nuclear ancestry into one of European, African, East Asian, South Asian or Other using genotype data from a set of SNPs selected as follows. We identified the 292,878 autosomal SNPs typed by three widely used Illumina genotyping arrays (HumanCoreExome-12v1.1, HumanCoreExome-24v1.0 and HumanOmni2.5-8v1.1). Presence of a SNP in all three arrays indicates that it can be reliably measured by microarray and, hence, that it is likely to be reliably called using genome sequencing as well. To further ensure this, we removed SNPs with a missing genotype in at least one individual or with an overall pass rate below 0.99. We also removed SNPs at genomic positions in which more than two distinct alleles had been observed in the 1000 Genomes Phase 3 dataset [26] or the NBR dataset to ensure that all genotypes could be coded unambiguously as a count of the number of copies of the unique alternative allele carried by the individual. We then removed SNPs with a MAF < 0.3 in our dataset. Finally, we pruned SNPs using PLINK v1.9 [21] so that all pairs of SNPs had an $r^2 < 0.2$. After filtering, 32,875 SNPs remained.

These well-measured, unlinked common SNPs were used to estimate relatedness in the non-Admixed individuals within the 1000 Genomes Phase 3 data as follows. First, we ran the snpgdsIBDKING function from the SNPRelate R package to compute an initial kinship matrix and identify a corresponding initial set of unrelated individuals [27, 28]. We then used PC-AiR [29] to perform a principal component analysis (PCA) on the standardised genotypes of these putatively unrelated individuals and project the standardised genotypes of the other individuals onto the fitted principal components (PCs). The PC-AiR object was then passed to the PC-Relate function to compute a kinship matrix that accounts for population structure as represented by the leading 20 PCs. Finally, the kinship matrix was passed to the PRIMUS [30] function to obtain a final set of pedigree relations and a final set of unrelated individuals on the basis that pairs of individuals with a kinship coefficient > 0.09 are related.

We partitioned these unrelated individuals as non-Finnish Europeans, Finns, Africans, South Asians and East Asians using their 1000 Genomes population code annotations. Within each element of this partition, we modelled the score vectors of the leading five PCs by a multivariate normal distribution and estimated the corresponding mean vectors and covariance matrices. Subsequently, we projected the genotypes from each of the 13,037 samples onto the vector space spanned by the leading five components of the 1000 Genomes PCAs (**Extended Data Figure 3h**). We computed the likelihood of the projected data under the five multivariate normal models estimated from the 1000 Genomes scores and labelled the individual with the population corresponding to the model that yielded the highest likelihood, provided it was greater than 10^7. Otherwise, the individual was labelled "Other" (**Extended Data Figure 3i**). The resource is predominantly European. However approximately 7% of enrolled individuals are South Asian, which is a group with limited whole genome reference data available through other public resources such as gnomAD (**Extended Data Figure 3j**).

To compute the relatedness amongst the 13,037 individuals in our dataset, we merged the genotypes with the 1000 Genomes genotypes and followed the procedure described above to compute relatedness a second time. As the 1000 Genomes collection is genetically diverse by design, combining the two datasets ensured that a greater amount of genetic diversity was accounted for by the PCA than if only the new sequence dataset had been used. The PC-Relate estimates of the genome proportions shared by zero, one and two chromosomes between two individuals were used to categorise pairwise relations and to identify a maximal set of 10,259 unrelated individuals (**Supplementary Table 1: Domain metrics**). A set of 9,244 pedigree networks ranging from singletons to a network of 15 were computed (**Figure 1d**).

Variant normalisation and loading of variants into HBase

Variants in the gVCF files were processed using the bcftools norm command with the -cs option, which left aligns and nomalises indels and sets/fixes incorrect or missing reference alleles [31]. The variants were then transformed from gVCF format into Google *proto* format [32] and stored as binary *proto* objects spanning 1000bp of contiguous sequence in the reference genome using OpenCGA [33]. During the transformation, the anchoring reference base of insertions and deletions were removed. The objects of each sample were then copied from the respective *proto* file to the HBase [34] *Archive* table using the 'PUT' operation of the HBase application programming interface. The resulting *Archive* table held the *proto* objects by sample in columns and grouped by position in rows. Header information from the original gVCF files was stored for each sample. Any multiple variant calls overlapping the same position within an individual were resolved by only retaining the variant with the highest genotype quality. The samples with each of the different genotype calls and whether these calls had a PASS filter value were recorded for different genomic positions in the HBase *Allele count* table using the 'APPEND' operation. Homozygous reference and PASS values were the most frequent and were inferred instead of being stored. Only the variants are then transferred from the HBase *Allele count* table to the HBase *Analysis* table.

SNV and indel annotations

All variants resulting from the merging across 13,187 samples require annotation with cohort summary statistics as well as deleteriousness and conservation scores. Cohort summary statistics including allele count, allele number, genotype count, minor allele frequency, minor genotype frequency, HWE, call rate, pass rate and overall pass rate (described below) were calculated. We used the HTSJDK [35] implementation to obtain HWE statistics. Subsequently, the variants stored in the HBase *Analysis table* were annotated using the RESTful annotation service provided by CellBase [36] in order to add consequence types, HGVS notation, deleteriousness scores like the combined annotation dependent depletion (CADD) score [37], SIFT [38] and PolyPhen-2 [39], and conservation scores like genomic evolutionary rate profiling (GERP) [40], PhastCons [41] and PhyloP [42]. Variants have then been exported to AVRO file format. These files have then been complemented with additional annotation from external reference datasets, including 1000 Genomes Project Phase 3 [43], UK10K project (version 2016-05) [44], the ExAC / gnomAD project (version r2.0.2) [45], Trans-Omics for Precision Medicine Program (TOPMed, freeze5) [46] and the Human Genome Mutation Database (HGMD, version PRO 2018.1) [47] using Apache Spark [48]. Since the TOPMed data is only available on genome reference GRCh38, these data were mapped back onto GRCh37 using CrossMap [49] before variant annotation.

SNV and indel quality control

At each position in the reference genome, and for each sample, the HBase database holds a record of whether the alignments at that position were such that variants could be called. Let the call rate be the proportion of samples in which calling was possible. In samples in which the genotype could be called, the genotype calling could either pass or fail QC. Let the pass rate be the number of samples with a passing genotype out of all the samples for which calling was possible. We defined the overall pass rate (OPR) as the call rate multiplied by the pass rate. Only female and male samples were used to set the OPRs for variants on the X and Y chromosome, respectively.

We evaluated the distribution of *p*-values under the null hypothesis of HWE for common SNPs and indels (MAF > 5%) amongst 8,511 unrelated Europeans within different tranches of OPR. Variants with OPR >= 0.99 were found to have genotype proportions consistent with HWE over a wide range of MAFs for SNVs, short insertions and short deletions (**Extended Data Figures 4a, 4b and 4c**). At 0.98 > OPR < 0.99, we observed a doubling of the number of *P*-values < 0.05 relative to the number expected under the null, approximately uniformly across MAF bins, suggesting that deviation from HWE was present in a small proportion of variants in this tranche of OPR. We thus applied a threshold of 0.99 for the release of a highly specific set of variants through the variant browser (described below), but used a threshold of 0.98 for rare variant analyses, where maintaining high sensitivity is of critical importance.

Concordance in duplicates and twins

The 125 duplicated samples and 22 monozygous twins can be used to estimate the reproducibility of sequencing results. From all resolved duplicates and twins, we selected pairs where both samples were sequenced with the same read length. The triplicate was not included.

For each sample we selected all autosomal SNVs and small insertions and deletions with OPR >= 0.99 and biallelic in this sample. **Extended Data Figure 4d** illustrates all possible combinations of genotypes in a pair of samples. If we denote with c and d numbers of variants in the cells corresponding to concordant and discordant calls, the mutual non-reference concordance can be defined as $(\Sigma c / (\Sigma c + \Sigma d)) * 100\%$, the distribution of this value is shown in **Extended Data Figure 4e**.

The probability of getting a heterozygous variant in sample 2 given it is heterozygous in sample 1 can be defined as $c_{11} / (d_{10} + c_{11} + d_{12} + d_{13})$, the same for homozygous variants: $c_{22} / (d_{20} + d_{21} + c_{22} + d_{23})$ (**Extended Data Figures 4f and 4g**). Assignment of which sample is sample 1 and which is sample 2 is random, hence for each pair of duplicates/twins there are two probabilities of each type. In the plots in **Extended Data Figures 4f and 4g** the lowest valued probability is shown in red and the highest in blue.

Overview of SNVs and indels in the dataset

We identified 172,005,610 variants with an OPR > 0.99, of which 157,411,228 (91.5%) were SNVs and 14,594,382 (8.5%) were indels. We annotated them with Sequence Ontology (SO) terms based on their predicted consequences with respect to transcripts in Ensembl 75. Thus, for each variant, multiple consequences were recorded. However, for the purposes of classification, we selected one primary consequence per variant as follows. First, the transcripts with the most relevant biotype (e.g. protein coding) were chosen [50]. Second, the transcripts with the best source of curation were retained (ranked by HGNC, VEGA, Ensembl, UniProt and miRBase/RFAM). Third, the consequences with the greatest Ensembl-summarised impacts were retained [51]; the most impactful being HIGH, MODERATE, LOW and MODIFIER, in that order). Fourth, consequences were preferred based on the order provided by Ensembl [51]. Fifth, transcripts which are canonical according to Ensembl were preferred over other transcripts. Finally, the transcript with the longest coding/cDNA sequence length was chosen.

Following these rules, we were able to identify a single primary consequence for each variant. Overall, 61.7% of the variants were labelled genic (SO:0001564, gene_variant), while the remainder were labelled intergenic (SO:0001628, intergenic_variant), regulatory (SO:0001566, regulatory_variant) or other. 80.0% of the genic variants were intronic (SO:0001627, intron_variant) and only 4.4% were exonic (SO:0001791, exon_variant) or affected a splice site (SO:0001568, splicing_variant). The ratio of coding (SO:0001968, coding_transcript_variant) to UTR (SO:0001622, UTR_variant) variants was 45:55 for SNVs and 21:79 for indels. The vast majority (64.2%) of the 1,853,212 coding SNVs were missense (SO:0001583, missense_variant) while 33.7% were synonymous (SO:0001819, synonymous_variant) and only 38,886 SNVs (2.1% of coding SNVs) led to a start loss (SO:0002012, start_lost), stop loss (SO:0001578, stop_lost) or stop gain (SO:0001587, stop_gained). The majority (62.1%) of the 67,746 coding indels introduced a frameshift (SO:0000865, frameshift), while 34.5% introduced an inframe insertion or deletion (SO:0001817, inframe) and only 2,287 indels (3.4% of coding indels) led to a start loss (SO:0002012, start_lost), stop loss (SO:0001578, stop_lost) or stop gain (SO:0001587, stop_gained) (**Extended Data Figure 5**).

48.6% and 59.2% of the SNVs and indels, respectively, were absent from the following databases: 1000 Genomes, UK10K, gnomAD, TOPMed (lifted back to GRCh37) and HGMD. 54.8% of the variants were observed in only one family (i.e. they had a minor allele count of 1 amongst unrelated individuals), of which 82.6% were novel. The proportion of variants observed in two families which were novel was 36.7%. Of the variants observed in more than two families, very few (1.90% of SNVs and 9.01% of indels) were novel (**Figure 1e**). Thus common variants are well represented in genetic databases but the vast majority of genetic variants are very rare and are mostly absent from the available databases.

Inside coding regions, indels that introduce frameshifts are, on average, selected against due to their deleterious effects on reproductive fitness [52], and our analysis corroborates this notion (**Extended Data Figure 3k**). In contrast, in non-coding regions of the genome, indels lengths tend to be a multiple of two bases (**Extended Data Figure 3l**). We have made a similar observation in the gnomAD dataset (data not shown) to ensure this is not a technical artefact specific to our bioinformatics pipeline. The pattern is specific to repetitive regions of the genome (**Extended Data Figure 3m**), suggesting this pattern can in part be attributed to tandem duplications.

Integrative Variant Analysis (IVA) web application

The SNVs and indels have been loaded into the IVA web application [53]. The data can be browsed freely online at http://bioinfo.hpc.cam.ac.uk/web-apps/nihrbr.

Large deletion calling and quality control

Two methods were used to call deletions larger than 50bp in each sample: Manta version 0.23.1 [54], which calls deletions on the basis of split read alignments and larger-than-expected insert sizes of read pair alignments, and Canvas version 1.1.0.5 [55], which calls deletions on the basis of sustained reductions in coverage over a continuous stretch of the reference genome. Manta is optimised for calling deletions of 50bp–10Kb while Canvas is optimised for calling deletions > 10Kb.

As two algorithms were used to call deletions independently in each sample, deletion calls were merged across methods within samples if the reciprocal overlap between them was > 70%. Whenever deletions were merged, they were annotated with the Manta breakpoints as they are more precise than Canvas ones. After merging, each deletion called in each sample was assigned to one of the three QC categories “Fail”, “Possible” or “Confident,” as follows.

“Fail”: any of the following criteria are met:

- Length is > 50Mb,
- Heterozygous genotype on non-PAR of chromosome X in a male,
- Located on the mitochondrial or the Y chromosome,
- Called by Canvas only and CANVAS_QUAL < 10,
- Called by Manta only and MANTA_FILTER is not “.”.

“Confident”: the call does not meet the “Fail” criteria and none of the following additional criteria are met:

- > 70% overlaps a flagged region (see below),
- Called by Canvas only and SNPS_DENSITY_HET_PASS > 0.5,
- Called by Canvas only in a sample in which the number of Canvas calls is greater than the mean plus five times the standard deviation (as such extreme values are likely due to non-uniformities in coverage leading to excessive false positive calls),
- Called by Canvas only in a sample with an excessive proportion of the genome deleted (greater than the 99.8% percentile across samples) if the number of deletions is greater than 50,
- Called by Manta only and MANTA_QUAL < 20th batch-specific percentile,
- Called by Manta only and MANTA_GQ < 20th batch-specific percentile,
- Called by Manta only and MANTA_IMPRECISE = “TRUE”,
- Called by Manta only and length > 1Mb,
- Called by Manta only and DUKE0_START = “TRUE” or DUKE0_STOP = “TRUE”,

- Called by Manta only and SNPS_N_HET_PASS > 0,
- Called by Manta only and heterozygous and READS_MEAN_MAPQ < 45.

The parameters in capitals above are generated by the Manta and Canvas software, except for DUKE0_START, DUKE0_STOP, SNPS_N_HET_PASS, SNPS_DENSITY_HET_PASS and READS_MEAN_MAPQ, which were generated using custom code. DUKE0_START and DUKE0_STOP flag deletions with start ± 10bp or end ± 10bp breakpoints within Duke non-unique region (see definition below); SNPS_N_HET_PASS is the number of heterozygous SNPs with OPR ≥ 0.99 within deletion boundaries; SNPS_DENSITY_HET_PASS is the ratio of SNPS_N_HET_PASS to the length of deletion in Kb; READS_MEAN_MAPQ is the mean mapping quality of reads within deletion boundaries. Flagged regions in the reference genome were defined as follows:

- Low mappability regions: DAC Blacklisted Regions (Encode, Accession: wgEncodeEH001432) and Human Mappability Blacklist (Encode, Accession: wgEncodeEH000322).
- Centromeres [56]
- Telomeres [56]
- IG loci (NCBI, Gene IDs: 50802, 3492, 3535)
- HLA loci (NCBI, Accession: NC_000006)
- Segmental duplications [57, 58]
- Duke non-unique regions (Encode, Accession: wgEncodeDukeMapabilityUniqueness35bp)

“Possible”: the deletion call does not meet either the “Fail” or the “Confident” criteria. Manual inspection of the reads using IGV of 100 deletion calls from the Possible and Confident deletion sets revealed that the false call rates were 24% and 10% respectively. To control the rate of false calls, whilst retaining calls in the Possible class which have a high chance of being real due to overlap with calls in the Confident set, any deletion in the Possible set which did not have a reciprocal overlap of at least 0.7 with any of the Confident deletions was removed.

In order to select a set of rare deletions for use in the downstream association analysis, the deletion calls required annotating with internal cohort allele frequencies. The deletion calls from all samples were partitioned into groups corresponding to distinct latent deletions. The partitioning was performed as follows. First, we used the ‘hclust’ function in R to construct a dendrogram of all deletion calls. One minus the reciprocal overlap was treated as the distance between each pair of calls, and the complete linkage option was used (i.e. the distance between two clusters corresponds to the maximum distance between two elements chosen from the two clusters). The dendrogram was cut at the position corresponding to a reciprocal overlap of 0.7 to create the partition, thus ensuring that all pairs of deletion calls within the same cluster had a reciprocal overlap of at least 0.7. To boost the number of real deletions retained, the partitioning algorithm was applied to the Confident deletions alone, and Possible deletions were subsequently assigned to the same groups as the Confident deletions with which they shared the highest reciprocal overlap. Applying this procedure resulted in 201,986 unique latent deletions.

Some sets of overlapping Manta deletion calls that were likely generated by the same underlying deletion had very different variances of left-hand and right-hand breakpoints across samples. This caused the partitioning procedure to fail to assign such calls to the same latent deletion. To address this, we employed the following remedy. First, we constructed a dendrogram using hierarchical clustering as described above on the 201,986 latent deletions. The start and end points of each latent deletion were estimated by taking the mean of the start points and the mean of the end points respectively across the Manta calls mapped to the latent deletion, or if there were no Manta calls mapped, taking the means across the Canvas calls. For each node in the dendrogram for which the mean reciprocal overlap of its descending latent deletions exceeded 0.5, we evaluated whether the spatial densities of start points or end points were significantly greater than expected by chance. We did this by performing a binomial test under the null that the rate of start positions or end positions in the interval was the same as that of the entire chromosome. If all *P*-values (one for each internal node in the dendrogram) were greater than a threshold, the procedure was halted, otherwise the deletions descending from the node with the lowest *P*-value were merged, and the procedure was then repeated. We selected a threshold of 0.05 which retained high specificity whilst removing a large number of highly overlapping clusters which appeared on manual inspection to represent the same latent deletions. Application of this merging step reduced the number of deletions by 24,436 deletions, resulting in a final 177,550 unique deletions being retained for analysis.

Retained deletion calls were then annotated with cohort allele frequencies based on the number of other deletions in their groups. We computed *P*-values under a null hypothesis of HWE for all common deletions as QC. The distribution of *P*-values was highly skewed (82% of deletions in unrelated Europeans with a cohort allele frequency of at least 0.05 had a *P*-value < 0.01) and we therefore concluded that the calling of common deletion genotypes was unreliable but retained rare deletions for downstream analysis (see below).

Clinical reporting

We performed diagnostic multi-disciplinary team (MDT) meetings for all domains except CNTRL, GEL, HCM, LHON and UKB. LHON focused on identifying modifier loci in nuclear DNA to explain incomplete penetrance of pathogenic mitochondrial variants, while HCM enrolled patients which had already been found negative in a screening for likely pathogenic and pathogenic variants in HCM diagnostic-grade genes. CNTRL and UKB are not ascertained to have a rare disease and the review of the GEL results is proceeding via the NHS Genomic Laboratory Hubs and the results will be reported in due course. No likely pathogenic or pathogenic variants were found for the 26 EDS samples analysed. Variant analyses and results described in the pertinent findings section exclude those six domains.

Gene lists and transcript selection

For each of the 15 rare disease domains (i.e. all domains except CNTRL, GEL and UKB) a gene list was generated by domain-specific experts. Genes were included in the lists if there was a high enough level of evidence for causality (more than 3 independent families reported, or 2 families but additional functional studies and/or a mouse model). The 2,478 gene/domain pairs, encompassing 2,073 unique genes across all domains, were manually curated and annotated with the relevant RefSeq and/or Ensembl transcript identifiers to support variant reporting. Transcripts were selected based on, by order of priority, community input, presence in the Locus Reference Genomic (LRG) resource [59] or designation as canonical in Ensembl (**Supplementary Table 2 (Diagnostic-grade genes)**). Genes in the 15 lists were submitted to LRG and 38.3% now have a curated transcript. These gene lists were introduced into the Sapientia^{TM} web application (Congenica Inc, Cambridge, UK) for use by the MDT to report causal variants to referring clinicians. There was significant overlap between the content of gene lists for the different domains (**Figure 2b**). During the course of the project, gene lists were updated and reversioned if new genes with an adequate level of evidence had been identified, and a final re-analysis of all samples was done on the most recent version of the gene lists. Hereafter, genes in the final lists are referred to as 'diagnostic-grade genes'.

Variant filtering

Variants (SNVs, indels) were shortlisted if (i) their MAF in control populations [22] was < 1/1,000 for putative novel causal variants and < 25/1,000 for variants listed as disease-causing in HGMD, (ii) their predicted impact according to the Variant Effect Predictor [60] was "HIGH" or "MODERATE" or if the consequences with respect to the designated transcript included one of "splice_region_variant" or "non_coding_transcript_exon_variant" if the variant was in a non-coding gene, (iii) the variant affected a gene relevant to the patient's disease. Variants with more than three alleles or a MAF >= 10% in the diseases cohort were discarded to, respectively, guard against errors in repetitive regions and remove potential systematic artefacts. The above filtering criteria were applied universally to all domains, except for ICP which adopted a higher MAF threshold of 3% for both novel and previously reported variants. The higher threshold accounted for causal variants being present in the male and non-child bearing female population. This strategy reduced the number of variants for review by the MDT from about 4 million per individual to fewer than 10, while confidently retaining known regulatory or moderately common pathogenic variants.

Web application for variant assessment

For each affected participant with prioritised variants, the variant calls, HPO-coded phenotype and the relevant metadata (unique study numbers; referring clinician and hospital; self-declared and genetically inferred gender, ancestry, relatedness, and consanguinity level) were transferred to Congenica for visualisation in the Sapientia^{TM} web application during MDT meetings. Sapientia^{TM} displays variant information such as predicted effect, location in the protein, MAFs in reference cohorts (e.g. ExAC, UK10K, and ESP [22, 44, 61], conserved regions, splice sites, and links to external resources (e.g. HGMD [47], ClinVar [62], OMIM [63] and PubMed [64], as well as showing patient data such as phenotype information in the form of HPO terms [14]. Sapientia^{TM} also allows annotation of each variant with its likely level of pathogenicity, the variant's contribution to the disease phenotype (partial, complete), and generation of tailored research reports for the referring clinicians.

Variant interpretation in MDT meetings

MDTs brought together experts from different hospitals across the UK and overseas, and typically consisted of an experienced clinician with domain-specific knowledge, a scientist with experience in clinical genomics, a clinical bioinformatician and a member of the reporting team. Assignment of the level of pathogenicity to variants followed the American College of Medical Genetics (ACMG) guidelines [65] and variants were marked in Sapientia^{TM} as pathogenic, likely pathogenic or of unknown significance (VUS). Only pathogenic and likely pathogenic ones were systematically reported and VUSs were reported at the MDT's discretion. As per REC-approved study protocol, secondary findings (e.g. pathogenic variants in *BRCA1* in patients not presenting with a relevant cancer phenotype) were not reported. Deletions involving domain-relevant diagnostic-grade genes were reviewed in MDT and reported if deemed likely pathogenic or pathogenic. Complex rearrangements were analysed for IRD and NDD [66]. The reports stated that clinical grade confirmation in accredited laboratory is recommended before feedback to the patient or modification of the clinical management. Although Sanger sequence confirmation data was not systematic collected on the >1000 reports issued, no failures to independently confirm the variants were reported. In addition, >200 variants were confirmed by Sanger sequence analysis or microarray internally and all variants reported were confirmed [4].

Sensitivity of WGS to detect HGMD variants

The mean depth of coverage by WGS was 36.07X across all samples. We compared the sensitivity obtained by WGS and by WES for variants catalogued as DM and DM? in HGMD Pro2018.1 by calculating the mean coverage for 1,000 representative male samples from our WGS data and the mean coverage provided by the ExAC dataset (**Extended Data Figures 6a and 6b**). A nominal cutoff of 20X (10X for variants on the non-PAR of the X chromosome) was used to distinguish between detected and non-detected to diagnostic standard. This analysis shows that 276 and 53 SNVs and indels, respectively, are not detected by either sequencing method and 75% and 64% of these variants, respectively, map to a small number of genes (*HBA1*, *HBA2*, *VWF* for SNVs; *HBA1*, *HBA2* for indels). There was a significant difference in sensitivity between WES and WGS, with 6,132 (5.89%) versus 139 (0.13%) of DM and DM? variants not being detected by WES and WGS, respectively. In keeping with this observation 96 (10%) of the 955 SNVs and indels reported by the MDT showed insufficient coverage in WES data and are thus likely not to have been detected by WES (**Extended Data Figure 6c**).

Overview of diagnostic results

In total 1,040 reports were generated by the MDTs listing 1,106 unique causal variants (733 SNVs, 263 indels, 104 large deletions, 6 other) with 299 (30.0%) of the SNVs and indels being absent from HGMD Pro2018.1 (**Supplementary Table 2 (SNV and indel list**, **Large deletion list)**). The reported variants were observed in 327 unique diagnostic-grade genes, with half of the reported variants being in 21 genes and a quarter in only three genes (**Figure 2d**).

The vast majority of causal variants in each domain were observed in fewer than five genes and the remaining variants were dispersed over large numbers of the remaining diagnostic-grade genes (**Extended Data Figure 7**). Interestingly, 21 causal variants were reported for diagnostic-grade genes shared between two domains (**Extended Data Figure 7c**), showing the pleiotropy of the phenotypic consequences of variants in the same gene. Two BPD cases with thrombocytopenia, platelet function abnormality accompanied by bleeding and a single NDD case with brachycephaly, microcephaly and global developmental delay, carried likely pathogenic missense variants in the phosphatase encoded by *PTPN11*, a gene known to harbour variants causal of Noonan syndrome [67, 68].

Altogether, reports with likely pathogenic and pathogenic variants were issued for 1,040 affected individuals, with diagnostic yields ranging between 1.6% for PMG and 53.9% for IRD (**Figure 2c**). The wide range in yield can be attributed to variation in the extent of genetic screening before enrolment (see Enrolment section), variation in the genetic architecture of disease and variation in the depth of knowledge of the genetic aetiologies of diseases. For example, the high yield for the IRD domain, where more than half of the affected individuals received a conclusive report, was achieved thanks to detailed phenotyping by retinal imaging, a lack of genetic pre-screening and a predominantly autosomal recessive mode of inheritance. In BPD, extensive phenotypic pre-screening was performed (e.g. full haemostasis testing). However, certain phenotypic traits, such as bleeding, are strongly influenced by environmental exposures (e.g. trauma, including surgery), and the genetic architecture is diverse, resulting in an overall yield of 12.7%. The reasons for the low diagnostic yield for PMG remains unclear, but there is emerging evidence that PMG may have a polygenic basis in most cases (manuscript in submission).

Six MPMT cases, two SMD cases and one NDD case carried a protein-truncating variant in *NF1* causative of type 1 neurofibromatosis, manifesting variously as neoplasm of the nervous system or small intestine (MPMT), dilatation and neurofibromas (NDD), and failure to thrive, broad philtrum, xanthomatosis, cafe-au-lait spot, hypotelorism, juvenile splenomegaly, enlarged kidney, hepatomegaly, monocytosis, myelomonocytic leukemia, and thrombocytopenia, anaemia (SMD). Conversely, some patients had phenotypes caused by variants in several genes. For example, the seven patients with a causal variant in *NMNAT1* or *RPE65* all have retinal dystrophy, as expected, but two also have intellectual disability likely caused by other variants.

The calling of deletions from WES results with adequate sensitivity and specificity remains challenging, particularly for short deletions encompassing only one exon or part of an exon and deletions of exons with poor coverage [69]. In contrast, WGS allows improved calling of large deletions. From the 177,550 unique deletions being retained for analysis, and after exclusion of those with a cohort frequency above 1 in 4,000 in our cohort of unrelated individuals and deletions that did not overlap diagnostic-grade genes for the case's domain, 524 deletions remained for visual review of reads in IGV. Nearly 85% (n = 444) unique deletions were deemed confirmed or likely to be real after this visual review and were presented to the relevant MDT. Of these, 23.42% (n = 104) were labelled as likely pathogenic or pathogenic (**Supplementary Table 2 (Large deletion list)**). The length of the 104 unique reported deletions ranged between

203bp and 16.80Mb (mean 786.33Kb; median 15.91Kb) and nearly half (n = 43) were absent from ClinVar.

A total of 28 heterozygous deletions encompassing autosomal recessive genes were returned to the referring clinician, with 25 of them reported alongside a SNV or indel on the other haplotype. For example, one SRNS case harboured a rare heterozygous 14.55Kb deletion that removed the last 7 exons of *NPHS1*, which, in conjunction with a heterozygous splice donor variant on the other haplotype, explains the patient's phenotype. This is the first report of a large deletion in *NPHS1* being causal of SNRS.

Because the project focussed on rare diseases, we observed an enrichment of probands from consanguineous families. When we reviewed the reported deletions, 15 were found to be present in homozygous state and nine of these were of probands born to consanguineous parents. For one case the apparent homozygosity of the deleted region was upon visual inspection of the reads in IGV rejected because it concerned an overlap between two unique heterozygous deletions originating from the parents. In the remaining five cases, the rare homozygous deletions were confirmed on inspection of the reads in IGV and by additional genotyping and uniparental disomy was excluded as a possible cause. Overall, 103 of the 1,040 reports issued (9.9%) included a large deletion, confirming the clinical importance of this category of variants in the pertinent finding analysis, and the value of using accurate deletion calling algorithms. In addition, for the IRD and NDD domains, we performed an additional analysis for complex structural variants (cxSV) and identified two cases harbouring a cxSV in a relevant diagnostic-grade gene [66].

The penetrance of causal genetic variants differed by domain. In general, causal variants in diagnostic-grade genes for IRD, NDD, BPD and SMD have high penetrance (conditional on a second causal allele being present on the other haplotype in the case of autosomal recessive diseases). In contrast, causal variants in the diagnostic-grade genes for ICP (*ABCB4*, *ABCB11*) result in cholestasis only when a woman becomes pregnant or takes specific drugs such as the combined oral contraceptive pill [70]. These variants are therefore relatively common in the unaffected population: in ICP, variants with a MAF as high as 1/270 in unaffected individuals have been previously reported as likely pathogenic [71] and we observed these variants in a large portion of the participants in the ICP domain. Causal rare variants in *BMPR2*, which are present in approximately 80% of patients with familial PAH and in 20% of patients with idiopathic disease, have been reported to have a penetrance as low as 14% in males [8, 72] and environmental triggers are considered important in precipitating this severe disorder. In NPD, penetrance is also modulated by environmental exposure for instance low temperature in the case of non-freezing cold injury. For the SRNS domain, there was extensive genetic pre-screening by whole exome sequencing and cases with likely pathogenic and pathogenic variants in known diagnostic-grade genes [73, 74, 75] have not been enrolled. This is one important reason for the relatively low diagnostic yield and new diagnostic-grade genes remain to be discovered for SRNS. It is also likely that there is a polygenic contribution to this condition.

Upload of reported variants to public repositories

Reported alleles and their clinical interpretation have been deposited with ClinVar, accession will be available before publication.

Reported alleles have been deposited with DECIPHER, accession will be available before publication.

Reported variants informing changes in clinical management

We did not systematically capture information whether the MDT-reported results led to changes in clinical management or the repurposing of existing drugs. There have however been some striking examples of how the latter has changed the care path of patients enrolled in this project. First, 27 patients with early-onset dystonia were shown to carry causal variants in the histone methyltransferase gene *KMT2B*, and many have been treated successfully by deep brain stimulation [76]. Second, gain-of function variants in *DIAPH1* cause macrothrombocytopenia and deafness [77]. We have recently shown that the treatment of such a patient with Eltrombopag, a FDA-approved drug for the treatment of autoimmune thrombocytopenia, increases the platelet count to a safe level in the perioperative setting [78], thereby reducing the need to use donor platelet concentrates. Finally, we identified a pedigree with a p.E527K gain-of-function variant in the kinase SRC resulting in juvenile myelofibrosis and severe thrombocytopenia, further complicated by osteoporosis. After the discovery of the pathogenic effect of this variant, the proband (case 27 of the published pedigree) was successfully cured of her thrombocytopenia by an allogeneic haematopoietic stem cell (HSC) transplant from her HLA compatible sister, who was negative for the causal variant [79].

There are also many examples of how the reported variants have led to the better stratification of patient care, including the frequency of clinic visits. We showed that haploinsufficiency of *NFKB1* is the most frequent cause of primary immune deficiency, with patients having recurrent and severe infections accompanied by autoimmunity and unexplained splenomegaly and an increased risk of oncological manifestations. These new findings have a direct impact on the care of this genetically defined category of PID patients [10]. Similarly, we identified 27 BPD cases with isolated thrombocytopenia caused by variants in *ANKRD26*, *ETV6* or *RUNX1*. These genes encode DNA-binding proteins and the identified variants are associated with an increased risk for haematological cancers, which is particularly increased for patients with variants in *ETV6* [80, 81] and *RUNX1* [82]. Hence frequent follow-up clinic visits are warranted and allogeneic HSC transplants needs to be considered as it provides an option for cure. In contrast, the 19 cases with thrombocytopenia due to variants in *ACTN1, CYCS* or *TUBB1* could be reassured of the benign nature of their condition [83] and such patients do not require regular follow-up but haematology consultation is required at times of haemostatic challenges (e.g. childbirth, surgical procedures, including dental ones). Finally the identification of four new genes (*ATP13A3, AQP1, GDF2, SOX17*) for PAH has led to an improved diagnostic yield for this severe condition and the genetic findings have also confirmed that mutations in *BMPR2* [84] and *EIF2AK4* carry a worse prognosis. In particular, patients carrying causal variants in *EIF2AK4* should be referred for early consideration for lung transplant [7].

Structural variants in patients with null protein phenotypes

Most of the patients enrolled to the BPD domain had comprehensive intermediate phenotype data relating to haemostasis, including measurements of coagulation factor levels (Factor I (fibrinogen; *FGA, FGB, FGG*), II (*F2*), VII (*F7*), IX (*F9*), XI (*F11*), von Willebrand Factor (*VWF*)), platelet receptor expression levels (e.g. of the fibrinogen receptor (GPIIbIIIa; *ITGA2B*/*ITGB3*) and red cell Rh grouping information. These data allows us to query unexplained cases who had a complete or near absence of one of these proteins, which we refer to as having 'null phenotypes'. Genetic defects relating to null phenotypes tend to have recessive and X-linked modes of inheritance for genes on the autosomes and chromosome X (*F9*), respectively. We identified 8 unsolved cases with null phenotypes for coagulation factors, 6 unsolved cases of possible Glanzmann's thrombasthenia (determined as cases having impaired aggregation in response to at least two of five agonists (arachidonic acid, collagen, epinephrine, thromboxane analogue, TRAP), but normal agglutination with ristocetin, and one patient with a mild bleeding disorder accompanied by an unexplained haemolytic anaemia lacking all common Rh groups from her red cells.

Three of the 15 cases had a rare coding variant on one haplotype of the relevant gene but the other haplotype appeared to be wild-type according to the standard variant calls. Through visual inspection of the reads in IGV, we were able to resolve two of these three cases.

The first of these had Glanzmann's thrombasthenia and carried a variant in *ITGB3* which encodes a premature stop at amino acid position 242. Visual inspection of the gene body revealed an excess of improperly mapped reads in intron 9, primarily due to alignment of paired-end reads facing away from, rather than towards, each other (**Extended Data Figure 8a**). We counted the number of improperly read pairs in the region in our collection after excluding the GEL domain participants (as we do not hold their phenotypic data), and found this case to have the greatest number of improperly mapped reads out of 6,656 individuals whose samples were sequenced using 150bp reads (**Extended Data Figure 8b**). We then analysed a further 304 in-house WGS samples sequenced in the same manner after the main data freeze and identified a further two individuals with an excess, primarily due to read mates aligning to different chromosomes (**Extended Data Figure 8c,d**). These two related individuals were from another Glanzmann's thrombasthenia case with his (the proband) platelets being devoid of GPIIbIIIa and the mother who showed a ~50% reduced expression of GPIIbIIIa (data not shown). The Glanzmann's case also carried a rare missense paternally inherited variant encoding a change from threonine to proline at amino acid position 456 of *ITGB3*. Sanger sequencing of the *ITGB3* cDNA obtained by reverse transcription of the RNA from the proband's platelets showed exclusive expression of the proline encoding allele and provided no evidence of an alternatively spliced transcript (data not shown). Oxford Nanopore-based sequencing of long-range PCR-amplified target DNA was performed as previously described [66] with the aim to resolve the genetic architecture of intron 9. The flow cell ran for 3 hours, and the mean coverage was 863,986X**.** We observed an insertion of a SVA (Alu, SINE-VNTR-Alu) retrotransposon (RE) element of 2,270bp at position chr17:45369041, predicted to induce nonsense-mediated decay. SVA elements are present throughout the reference genome, explaining the short read mates aligning to various chromosomes (**Extended Data Figure 8e**). To date, roughly 130 pathogenic variants caused by retrotransposon elements have been documented in the literature [85], but none of them have been implicated in Glanzmann's thrombasthenia. Here we report a first example and demonstrate the utility of long-range sequencing reads in resolving causal variants of a complex nature suggested by short-read WGS analysis.

The red cells of the second resolved case serotyped negative for all common antigens of the Rh blood group system. Upon further investigation the red cells were shown to lack both the RHD and RHCE proteins and also the Rh-associated glycoprotein RHAG. This very unusual Rh-null red cell phenotype is known to be cause a syndrome characterised by chronic haemolytic anaemia of varying severity [86]. The most frequent genetic explanation for this rare condition are loss-of-function variants for both *RHAG* haplotypes, because the RHAG protein is essential for the expression of the RHD and RHCE proteins on the red cell membrane [87]. This Rh-null case was shown to have a known causal loss-of function splice donor acceptor variant in intron 2 of *RHAG* [88], but a second event on the alternate haplotype was not identified by automated variant calling. Upon visual inspection of the reads of the *RHAG* locus, we identified the second event being a heterozygous tandem duplication spanning exons 2–7.

Genetic association between rare variants and rare diseases

Statistical approach

We used the BeviMed statistical method [89] to identify genetic associations with rare diseases in our dataset. Each run of BeviMed requires the definition of a set of cases and controls, all of which should be unrelated with each other, and a set of rare variants to include in the inference. To achieve adequate power, the cases should be chosen such that they potentially share a common genetic aetiology (e.g. because the phenotypes are similar) and the rare variants should be chosen such that they potentially share a mechanism of action on phenotype (e.g. because they are predicted to have a similar effect on a particular gene product). BeviMed computes posterior probabilities of no association, dominant association and recessive association and, conditional on dominant or recessive association, it computes the posterior probability that each variant is pathogenic. We can impose a prior correlation structure on the pathogenicity of the variants that reflects competing hypotheses as to which class of variant is responsible for disease. These classifications typically group variants by their predicted consequences. The class of variant responsible can then be inferred by BeviMed, thereby suggesting a particular mechanism of disease.

Case/control groupings using phenotypic tags

A set of phenotypic 'tags' were defined for each domain that determined case/control groupings for BeviMed. Cases shared a particular tag if their phenotypic characteristics were considered compatible with a shared genetic aetiology of disease. Tags could be set using logical rules applied to the HPO terms and other data or they could be set manually. The full set of tags and corresponding numbers of cases and controls can be found in **Supplementary Table 3 (Phenotypic Tags)**. Given a particular tag and a set of rare variants, the corresponding case/control groupings were obtained as follows:

1. Cases: unrelated (i.e. at most one per family) affected individuals with the tag who have not been explained by other variants.

2. Controls: members of the maximal set of unrelated individuals who do not have the tag and who are not related to any of the cases.

Selection of variants affecting transcript sequences

For each gene, we selected SNVs/indels for inclusion in the BeviMed analyses as follows. First, let us define $PMAF_X$ for a given variant as the probability that the minor allele count is at least the observed minor allele count, given that MAF = 1/X. Only variants with an internal MAF amongst unrelated individuals less than 0.002 were used. Additionally, variants had to have a $PMAF_{500}$ less than 0.05 in European and non-European probands and in European and non-European members of the maximal unrelated set to avoid bias in the association analyses. SNVs/indels were retrieved from the HBase variant database and were filtered as follows:

- Minimum OPR >= 0.98 for each of the three batches
- Unless a variant was in HGMD with class DM/DM?, it had to have $PMAF_Z$ < 0.05 in every gnomAD population (only gnomAD males were used to compute the $PMAF_z$ on variants in the non-PAR of X). We used Z =1,000 for recessive association analyses and z =10,000 for dominant association analyses.
- Variants in the non-PAR of X that appear only as heterozygotes in males were excluded.
- Multi-allelic variants for which the reference allele was the minor allele were excluded.
- Unless a variant was in HGMD with class DM/DM?, it had to have a CellBase-predicted consequence of type transcript_ablation, splice_acceptor_variant, splice_donor_variant, stop_gained, frameshift_variant, stop_lost, start_lost, transcript_amplification, inframe_insertion, inframe_deletion, missense_variant, protein_altering_variant, regulatory_region_ablation, non_coding_transcript_exon_variant or 5_prime_UTR_variant on a transcript of type lincRNA, miRNA, misc_RNA, Mt_rRNA, Mt_tRNA, protein_coding, rRNA, snoRNA, snRNA, TR_C_gene, TR_D_gene, TR_J_gene, TR_V_gene, IG_C_gene, IG_D_gene, IG_J_gene or IG_V_gene ("retained biotypes").

Large deletions with a $PMAF_{500}$ less than 0.05 in European and non-European probands, and in European and non-European members of the maximal unrelated set were selected for use in the BeviMed analysis.

The worst predicted consequence of a variant with respect to a particular gene was annotated as being “5’ UTR” if there were no consequences with HIGH or MODERATE impact and at least one transcript was annotated with the 5_prime_UTR_variant consequence, respectively. The worst predicted consequence was annotated as being “moderate” if at least one transcript was annotated with a MODERATE or HIGH impact or the non_coding_transcript_exon_variant consequence. The worst predicted consequence was annotated as being “high-impact” if at least one transcript was annotated with a HIGH impact or the variant was a large deletion that overlapped an exon of a transcript with a retained biotype. Thus, all variants in the high-impact class were also in the moderate class, while variants in the 5’ UTR class were not present in any other classes. The rationale for this is that a missense variant can be loss of function and thus have an equivalent effect as a truncating variant such as a frameshift. The prior probabilities of the models that included each of the four groupings of variants were 0.01, 0.495, and 0.495 for the “5’ UTR”, “moderate” and “high-impact” variant classes, respectively. All priors in the BeviMed model were left as default, including a prior probability of association across all association models of 0.01.

BeviMed results for genes

BeviMed was run for each tag/gene pair. For each pair, the total posterior probability of association, summing over all modes of inheritance and classes of variants (5’ UTR, moderate and high-impact) was recorded. A total posterior probability greater than 0.75 was considered good evidence of a causal relationship. The BeviMed model assumes that cases and controls are drawn from the same population and, consequently, that non-causal variants have the same MAFs in cases and controls. Using external data from gnomAD, variants with a MAF likely to be above a given threshold in any gnomAD-defined population were excluded (see above). However, due to the small sample sizes for non-European populations in gnomAD, modelled variants observed in non-European study participants were more likely not to be rare in their respective populations than modelled variants observed in European study participants. If these variants were carried by non-Europeans with the same tag, a false positive association could result. In order to guard against this risk, we removed variants observed exclusively in at least two people of non-European ancestry and re-ran each analysis. The original tag/gene pairs for which the total posterior probability of association in these subsequent analyses dropped below 0.1 were removed. The results that remained are shown in **Figure 3**.

Rare variants associated with phenotypic extremes in the UK Biobank cohort

A genomewide association study (GWAS) of FBC traits measured in approximately 173,000 European ancestry participants in the UK Biobank cohort (n~133,000) and the INTERVAL trial (n~40,000) [90] previously identified 582 genetic variants independently associated with quantitative properties of mature red blood cells [91]. Most of these variants were common, with only 40 having an in sample MAF lower than 1%. However, the GWAS design had limited power to identify associations with rare variants due to the imprecision of genotype imputation (Supplementary Materials, Appendix 2). In order to identify rare variants associated with full blood count measured properties of red cells, we attempted to derive a univariate quantitative score with high rare-variant heritability by using the previously reported GWAS associations of 65 variants modulating properties of red cells of with MAF<1% as a model for the effect of rare variants on mature red cell parameters (**Figure 4a**). 384 individuals were selected for WGS from each tail of the score, of which 383 from the left tail and 381 from the right tail passed quality control after successfully sequencing (**Figure 4b**). The left tail of the score (BeviMed tag “Left”) represents individuals with lower than average RBC# and higher than average MCV, while the right tail (BeviMed tag “Right”) represents individuals with higher than average RBC# and lower than average MCV (**Figure 4c**). The participants from each tail that passed quality control were treated as a rare-disease case group in a BeviMed analysis. The inferred posterior probabilities for the genes with the strongest evidence for association (posterior probability greater than 0.4) are displayed in **Figure 4d**. The strength of evidence (respective posterior probabilities 0.997 and 0.575) that rare variants in *HBB* and *TFRC* are associated with the low MCV/high RBC# phenotype (the right tail of the score) is entirely consistent with biological knowledge (**Supplementary Table 3 (BeviMed Association UK Biobank)**) and these associations can be considered ‘positive controls.’ Similarly, prior biological knowledge suggests that the association identified with the transcription factor *CUX1* is very plausible while the associations identified with *ALG1*, *ZNF407* are plausible. The weight of evidence from a detailed examination of the alleles implicated in the other genes exhibiting a BeviMed posterior probability of association > 0.4, together with the expression profiles of the genes in blood cells and other tissue types is less compelling. In conclusion, our analysis demonstrates that genomic loci carrying rare alleles causing large deviations in quantitative traits can be identified by applying the BeviMed analysis approach. Since the extreme tails of any heritable quantitative trait are likely to be under negative selection, mutations in such genes may cause rare diseases or be otherwise clinically relevant. The forthcoming WGS of the full UK Biobank cohort offers a further opportunity to search for rare variant associations with other biomedically relevant quantitative intermediate traits, including blood cell traits not explored here.

Matchmaker Exchange

GeneMatcher [92], which is part of Matchmaker Exchange [93], was used to identify patients outside our collection with variants in candidate disease genes identified in our collection through a manual review process informed by the literature. Two syndromic BPD patients with platelet dense granule storage pool disease in addition to having an NDD had a match. The first patient with hypotonia, mental disability, epilepsy, uncontrolled movements and gastrointestinal problems matched with *SLC18A2* (coding for a serotonin transporter VMAT2) deficiency and a complete lack of serotonin storage in his platelet dense granules was detected (submitted). The second patient with psychomotor retardation and epilepsy, matched with *MADD* deficiency, a candidate gene for platelet function [94]. For both these patients, co-segregation analysis supported the presence of one recessive variant in each parent. Additionally, an NDD patient with intellectual disability, autistic features and seizures was linked via GeneMatcher to three unrelated patients with a similar phenotype, and a fourth patient was connected by personal correspondence [95]. In those five patients, the *WASF1* gene had been independently identified as a strong candidate because of features consistent with those of developmental-disorder-associated genes. It is constrained for loss-of-function variation in the ExAC Browser (pLi = 0.91) and is highly and specifically expressed in the adult human brain [96]. All variants are *de novo*, loss of function, and absent from 1000 Genomes, the ExAC and gnomAD data; interestingly, they cluster around the WASP-homology 2 (WH2) domain, included in the highly conserved C-terminal actin-binding WCA region regulating *WASF1*.

Regulome analysis

Ethics

Samples and information for generating open chromatin and histone modification data for activated CD4+ (aCD4) T-cells were collected with written and signed informed consent. The study was approved by the East of England – Cambridgeshire and Hertfordshire REC reference 05/Q0106/20. The other regulome data have already been released as part of previous studies and information about the consent under which these samples were collected can be found in the publicaitons related to these former studies.

Definition of cell type specific regulomes

As WGS provides variant calls across the entire genome, we sought to identify rare variants that exert their effect on phenotype through disruption of non-coding (possibly non-exonic) regulatory elements, such as enhancers. These elements have not been systematically assayed in past studies that used whole-exome or other targeted sequencing. Regulatory elements are largely specific to particular cell types. Thus, in order to focus the search for regulatory pathogenic variants, we first defined the coordinates of elements for different cell types implicated in the disease domains (**Supplementary Table 3 (Regulome Cell Data)**). We call each set of regulatory elements corresponding to a cell type a regulome. We defined regulomes for aCD4 cells, B cells (B), erythroblasts (EB), megakaryocytes (MK), monocytes (MON) and resting CD4+ T cells (rCD4).

For each cell type, we used open chromatin data (ATAC-seq) and histone modification data (H3K27ac) to identify regulatory elements using the RedPop method (see below). Additionally, for MK and EB, we had access to the following transcription factor (TF) ChIP-seq data, which were used to call peaks (see below) and supplement the regulomes: FLI1, GATA1, GATA2, MEIS1, RUNX1, TAL1 and CTCF for MK; GATA1, KLF1, NFE2 and TAL1 for EB; and CTCF for MON and B.

For each cell type, the regulome build process proceeded as follows:

1. Call RedPop regions using ATAC-seq/DNAse-seq and H3K27ac-seq data.
2. Call TF binding peaks using ChIP-seq data if available and obtain enrichment scores.
3. Discard TF regions with an enrichment score < 10 unless they overlap between at least two different TFs.
4. Collapse overlapping features to obtain a single genomic track.
5. Merge features within 100bp of each other.

Each regulome feature was assigned a gene label using either gene annotations from Ensembl (v75) or a compendium of previously published promoter capture Hi-C data (pcHi-C) [97] as follows:

1. Assign to a gene if the feature overlaps the gene or the region up to 10,000bp either side of the gene body.
2. Assign to a gene if the feature overlaps the gene's pcHi-C 'blind' spot. This region is defined by three *Hind*III restriction fragments, incorporating the capture fragment overlapping target gene TSS, and 5' and 3' adjacent fragments.
3. Assign to a gene if the feature overlaps a linked promoter interacting region identified using pcHi-C in the same cell type.

Regulatory element detection using patterns of peaks (RedPop)

We sought to identify regulatory elements in cell types of relevance to the rare disease domains in this project in order to search for possible pathogenic variants within them. It is important to locate these elements because they are much more likely to harbour pathogenic variants than other non-coding regions of the genome. Open chromatin data alone do not provide sufficiently high spatial resolution and transcription factor ChIP-seq restricts to regions bound by proteins targeted by specific antibodies. Open chromatin and histone modification (H3K27ac) sequencing data together, on the other hand, can be used to detect regulatory elements with high resolution and in an unbiased fashion. Open chromatin around binding sites typically results in a broad, low-resolution peak of elevated ATAC-seq/DNAse-seq coverage. The surrounding nucleosomes of a regulatory element are typically acetylated, leaving two peaks in H3K27ac coverage, spaced a few hundred bp apart. By combining the genomic coverage tracks of an open chromatin and an H3K27ac assay, regulatory elements can be detected with high precision. We developed an algorithm for regulatory element detection using patterns of peaks (RedPop) (manuscript in preparation) that utilises these patterns.

First, ATAC-seq/DNAse-seq and H3K27ac ChIP-seq reads were aligned to the genome using the BWA [98] aln command. Open chromatin peaks were then called using F-Seq [99]. Additionally, an open chromatin coverage track was generated, which was normalised by dividing by the mean coverage genome-wide and smoothed by binning consecutive segments of 40bp. Peaks were extended upstream and downstream symmetrically until their lengths were at least 3.2Kb and overlapping segments were subsequently merged. These merged segments were subsequently considered separately.

The covariance between the open chromatin track and the H3K27ac track was computed in 800bp sliding windows and subsequently smoothed by replacing each covariance value with the mean of the values in the surrounding 800bp. Local minima of the smoothed covariance were obtained as the positions for which the value was less than the values in the surrounding 160bp. Any local minima with a smoothed covariance less than -1 was recorded.

For each local minimum, the stretch nearest to it and any other stretches within 100bp of it, for which locally normalised open chromatin coverage exceeded the locally normalised H3K27ac coverage were recorded and expanded to 400bp, where locally normalised coverage at a position was given by the coverage divided by the mean coverage in the surrounding 800bp region. These stretches are retained as long as the open chromatin coverage track exceeded 47X (a default value obtained through a sensitivity/specificity study of data from MKs), otherwise they were discarded. The retained stretches were merged and recorded as the locations of regulatory elements.

Transcription factor binding peak calling

We applied the BLUEPRINT protocol for chromatin immunoprecipitation sequencing (ChIP-seq) data analysis. [100] Briefly, H3K27ac histone modification sequenced reads were mapped to human genome GRCh37 with BWA [98] aln method. Low-quality reads (-q 15), multi-mapped and duplicate reads were marked and removed with samtools and picard [101] respectively. ChIP enrichment was estimated with deepTools plotFingerprint [102] and samples were merged together to obtain one alignment file per cell type. H3K27ac peaks were called with MACS2 [103] with narrow option. For each cell-type the corresponding input file was used and downsampled according to ChIP size.

Identifying possibly causal deletions of elements regulating diagnostic-grade genes

In order to identify regulatory elements from the list described above that are involved in rare diseases, we examined rare deletions with a $PMAF_{500}$ less than 0.05 in European and non-European probands overlapping these elements. We further filtered these deletions to retain only those for which one of the putative target genes for an overlapping element is a diagnostic-grade gene associated with recessive disorders for the domain of the sample with the deletion, the cell type for which the element was called is relevant to the domain of the sample (**Supplementary Table 3**), and at least one of the following is true:

- the gene harbours a moderate impact rare allele (selected as described in the *Genetic association testing in genes* section) in the same sample,
- the deletion was called homozygous or hemizygous by either Manta or Canvas.

Our approach is illustrated in **Figure 5a**. Our filtering resulted in a list of only four deletions: a heterozygous deletion overlapping the 5' UTR region of the *ARPC1B* found in a PID patient who also carries a frameshift variant in the same gene (Thaventhiran et al, under review); a hemizygous deletion of a *GATA1* enhancer in case with thrombocytopenia, described below; a homozygous deletion of a CTCF binding site in the first intron of *LRBA* in a PID patient, described below; and a deletion which on manual inspection proved to be a technical error.

Deletion of GATA1 enhancer and HDAC6 open reading frame

Screening for deletions of MK-specific enhancers in BPD patients that are absent in the other rare disease cohorts identified a hemizygous 4108bp deletion (X:48,659,245-48,663,353) that comprised a potential MK enhancer and the four first exons of the histone deacetylase 6 (HDAC6) gene (Figure 5b, Extended Data Figure 9a) in a 9-year old boy with macrothrombocytopenia, bleeding symptoms and mild intellectual disability (ID) with an autism spectrum disorder (ASD) (Figure 5c). His parents are healthy though mild asymptomatic thrombocytopenia was also present in the mother (Figure 5c). The patient was hemizygous for the deletion while the mother was a heterozygous carrier (Figure 5d, Extended Data Figure 9b). Platelets of the propositus express no HDAC6 protein as the deletion removes part of this X-linked gene including the ATG start site (Figure 5e). In contrast, HDAC6 expression in platelets from the mother was comparable to expression levels in platelets from unrelated controls and the father. HDAC6 is the major deacetylase responsible for removing the acetyl group from Lys40 of α-tubulin, which is located in polymerized microtubules [104]. Indeed, absence of HDAC6 expression in platelets was accompanied with very high expression levels of acetylated α-tubulin while non-acetylated α-tubulin was expressed at similar levels as controls (Figure 5e). Platelets from the mother but not the father contain at least some platelets with hyperacetylated α-tubulin levels (Figure 5e). HDAC6 is extremely intolerant to LOF variants (pLI:1 and %HI:48.03) and the propositus is the only patient in our cohort with a hemizygous high impact variant. Figure 5b illustrates that the deleted region also contains binding sites for several transcription factors important for megakaryopoiesis [105]. In addition, this region overlaps with a recently identified regulatory element upstream of HDAC6 that was shown to strongly control GATA1 expression in megakaryoblastic K562 cells [106]. Interestingly, GATA1 protein was significantly decreased in platelets from the propositus and mother (Figures 5f and 5g). GATA1 is an important transcription factor that regulates platelet formation and hemizygous variants result in macrothrombocytopenia [107].

Platelets of the propositus have an increased Mean Platelet Volume (MPV) and Platelet Distribution Width (PDW) while for the mother only the PDW was increased (Figure 5c). Electron microscopy (EM) analysis of platelets from the propositus confirmed larger and rounder platelets with fewer alpha granules, very similar to GATA1 deficient platelets [108] (Extended Data Figure 9c, quantification shown in Extended Data Figures 9d and 9e). HDAC6 deficiency has never been described for humans while Hdac6 knockout mice have no gross defects [109] except for alternated emotional behaviour [110] and enhanced platelet spreading due to their hyper-acetylated microtubules while their bleeding tendency was not evaluated [111]. Structured illumination microscopy (SIM) analysis of acetylated tubulin in combination with F-actin (Extended Data Figure 9f) was performed for platelets under basal and activated conditions. Non-activated platelets of the propositus show disturbed marginal bands that are hyperacetylated (Extended Data Figure 9g). Quantification of platelet spreading on fibrinogen showed enhanced spreading for the propositus while platelets from the parents are similar to the control (Extended Data Figure 9h). Hdac6 knockout mice have normal platelet counts and exhibit normal megakaryopoiesis in contrast to in vitro studies using human MKs with HDAC6 depletion using shRNA or inhibition using the HDAC6 inhibitor Ricolinostat that resulted in a defective proplatelet formation [112]. Peripheral haematopoietic stem cells from the propositus, his parent and an unrelated control were differentiated to MKs and a significant reduction in proplatelet-forming MKs was observed for the propositus while a milder defect was also detected for the mother (Extended Data Figure 9i). All MKs from the patient and some cells from the mother show absent HDAC6 expression (Extended Data Figure 9j). In contrast to MKs treated with HDAC6 inhibitors, we detect obvious microtubule detects in MKs for the patient and mother (Extended Data Figure 9k). Proplatelet formation defects have also been observed in MKs from GATA1 deficient patients [108] and therefore it is difficult to distinguish between the contributions of GATA1 versus HDAC6 deficiency for the MK defect. In contrast, the platelet phenotypes observed in the propositus seem to combine a GATA1 defect with low GATA1 expression, macrothrombocytopenia and fewer alpha granules combined with hyper-acetylated microtubules resulting in enhanced platelet spreading due to HDAC6 deficiency.

Details on the functional analysis of the GATA1 enhancer/HDAC6 deletion

PCR and Sanger sequencing to validate the HDAC6 deletion. Genomic DNA was extracted from peripheral leukocytes. PCR was performed with the primers flanking the deletion (HDAC6-F: 5’-catcttcaagaggatcagagg and HDAC6-R: 5’- catagctagacactggtt), generating a PCR fragment of 358 bp when the deletion is present. Sanger sequencing of PCR fragments was perform using the same primer sets.

Antibodies. The following antibodies were used: rabbit HDAC6 (clone D2E5, Cell Signaling technology, Danvers, MA, USA), mouse anti-acetylated tubulin antibody (clone 6-11B-1, Sigma, St Louis, MO, USA), mouse anti-alpha-tubulin (A11126, Thermo Fisher Scientific, Waltham, MA, USA), rabbit VWF (Dako, Aligent Technologies, Leuven, BE), mouse CD63 and rat GATA1 N6

(Santa Cruz Biotechnology, Dallas, TX, USA), rabbit GATA1 (NF that was produced against recombinant N-terminal zinc finger, [113], rabbit GAPDH (14C10, Cell Signaling) and integrin beta3 (sc- 14009; Santa Cruz Biotechnology).

Electron and fluorescent microscopy of platelets and immunoblot analysis. Electron microscopy for platelets from the propositus was performed as described [107]. Immunostaining of resting and fibrinogen spread platelets was performed for platelets from the propositus, parents and an unrelated healthy control as previously described [114]. Platelet imaging was performed using a structured illumination microscope (SIM, Elyra S.1, Zeiss, Heidelberg, D.E). Images were analyzed with ZEN Black (Zeiss, Heidelberg, DE). Images were analyzed with ImageJ software (National Institutes of Health, Bethesda, Maryland, U.S.A) using the 'Analyze Particles' plugin for automated analysis. Total protein lysates were obtained from platelets as described [115]. Protein fractions were resolved by SDS–polyacrylamide gel electrophoresis, and blots were incubated with the indicated antibodies. Membranes were next incubated with HRP-conjugated secondary antibody, and staining was performed with the ECL detection reagent (Life Technologies). Chemiluminescent blots were imaged with the ChemiDoc MP imager, and the ImageLab software version 4.1 (Bio-Rad) was used for image acquisition.

Hematopoietic stem cell differentiation assay. CD34+ hematopoietic stem cells (HSC) were isolated by magnetic cell sorting (Miltenyi Biotec) from peripheral blood from the propositus, his parents and an unrelated control. The recovered (differentiation day 0) CD34+ stem cells were cultured in StemSpan SFEM medium with StemSpan CC100 ensuring strong expansion of HSC for 3 days (Stem Cell Technologies, Vancouver, C.A). Differentiation was next initiated by adding 50ng/ml thrombopoietin (TPO), 25ng/ml stem cell factor and 10ng/ml interleukin 1β (Peprotech, Rocky Hill, New Jersey, U.S.A). Proplatelet formation counting (on total differentiation day 12) as previously described [114, 79]. For immunostaining MK were seeded for 4 hours on fibrinogen-coated coverslips and stained cells were photographed at 63x magnification with a confocal microscope (AxioObserver.Z1, Zeiss, Heidelberg, D.E).

Homozygous deletion of CTCF binding sites in the first intron of LRBA

LRBA is a member of the gene family of BEACH-domain containing proteins and has been recently identified as a novel diagnostic-grade gene for the PID domain [116]. Homozygous coding mutations causing loss-of-function of *LRBA* are causal of a syndrome characterized by early onset hypogammaglobulinemia with autoimmunity. We noted an unresolved PID case who presented with a mild pancytopenia, characterised by mostly neutropenia and autoimmune haemolytic anaemia, occasionally complicated by periods of thrombocytopenia (**Figure 5h**) and with a homozygous deletion of a CTCF binding site in an element proximal to the *LRBA* promoter. The clinical features of this PID case are compatible with reduced LRBA function and it is thus plausible that the deleted CTCF-binding element is causally implicated in the pathologies observed in this patient.

Identifying possibly causal SNVs in elements regulating diagnostic-grade genes

In order to identify possibly causal non-coding SNVs, we examined rare SNVs with a CADD phred score > 20 which overlapped a regulatory element of a diagnostic-grade gene associated with a recessive disorder. Both the cell type in which the element was called and the sample had to be labelled with the same domain (**Supplementary Table 3**). Furthermore, the sample had to have a HIGH impact rare variant such as a deletion or premature stop in the body of the diagnostic-grade gene. Application of this procedure yielded two candidate SNVs in elements regulating *AP3B1* and *MPL*. The SNV in the *MPL* element was followed up for further analysis.

An SNV in the promoter of MPL, combined with deletion of exon 10 of MPL

MPL encodes the receptor for the megakaryocyte growth and development factor [117] thrombopoietin. Coding loss-of-function mutations on both alleles of *MPL* are causal of chronic amegakaryocytic thrombocytopenia (CAMT) [118]. CAMT can be categorised as type 1 or type 2 depending on the severity of the thrombocytopenia and ensuing bone marrow aplasia. The bioinformatic approach outlined above highlighted a thrombocytopenic 10-year-old male with a single exon heterozygous deletion of *MPL* (chr1:43,814,723-43,815,177) and a heterozygous SNV with a CADD score of 21.8 which is absent from gnomAD. The SNV lies in a strong MK-specific regulatory element (**Figure 5i**). Motif analysis using MatInspector [119] predicts binding to HIF1 of the wild type sequence (GGAC**G**TGGGGCT) through its very well-characterised recognition site “RCGTG”, but not of the mutant sequence (GGAC**A**TGGGGCT). The patient presented in his first months of life with a rash and a full blood count (FBC) analysis revealed a platelet count of 45 x 10^9/L. This low count was thought to be secondary to viral infection and no further investigations were performed at that time. At the age of 4 years, a routine FBC done during a consultation for his attention deficit hyperactivity disorder, which was assumed to be caused by a delivery-related trauma, showed the low platelet count to be chronic in nature. A bone marrow aspirate showed reduced numbers of megakaryocytes, adequate erythroid and plentiful myeloid precursor cells and no signs of myelofibrosis. A clinical diagnosis of a CAMT-like condition was made but could not be genetically confirmed because of a lack of a second coding variant with consequences in *MPL*. The mother, who carries the large deletion (**Extended Data Figure 10a**) and the father of the proband are healthy and have FBC results within normal ranges. The *MPL* regulatory SNV (chr1:43803414 G>A) is in trans of the large deletion because it is absent in the mother and therefore was inherited from the father or is a *de novo* variant. Measurement of the level of the MPL protein on the platelets from the proband and the mother by flow cytometry using a specific monoclonal antibody showed markedly reduced levels of reactivity for the proband compared to controls and the mother. In conclusion, absence of the MPL protein due to coding loss-of-function variants on both *MPL* alleles is causal of type 1 CAMT. The case reviewed above has a chronic thrombocytopenia but the other blood cell lineages seem unaffected. This clinical phenotype is compatible with the notion that the reduced level of MPL protein in this case is sufficient to prevent haematopoietic stem cell exhaustion, which is hallmark of type 1 CAMT.

Alternative variant datasets for versions 37 and 38 of the human reference genome

In order to migrate the 13,187 whole genomes to the human genome reference GRCh38, the Genalice high performance NGS secondary analysis suite [120] was deployed. The sequencing reads were extracted from the bam files delivered by Illumina and for comparison mapped against both assemblies, GRCh37 and GRCh38, using Genalice Map (v2.5) using default parameters [121]. The mapped reads were stored in a proprietary file format, called Genalice

Aligned Reads (GAR). SNVs and indels were subsequently called using the Genalice Population calling tool in single sample mode. The variants were collated together in a single Genalice Variant Map (GVM) with blocks of reference matching positions and quality metrics (i.e. calling quality, genotype likelihoods, etc.). All genomes were processed against either of the assemblies in less than 14 days (read mapping: 12 days; genotyping: 2 days) using 10 compute nodes (Intel(R) Xeon(R) Gold 6142 CPU, 32x 2.60GHz). Variants were then exported in AVRO and standard VCF file format for variant annotation and further downstream analysis.

Appendix

Appendix 1: Neuropathic pain disorders

Neuropathic pain arises as a consequence of a disease or lesion in the somatosensory nervous system [122]. A number of extreme neuropathic pain phenotypes, caused by rare high impact genetic mutations have recently been described [123]. Identification of such mutations has implications for diagnosis, genetic counselling and in some cases personalised treatment [124]. In broader terms such mutations help us understand the pathophysiology of neuropathic pain with implications for more common acquired neuropathic pain disorders, such as painful diabetic neuropathy [125]. Loss of sensation can be caused by inherited sensory neuropathies with sensory loss restricted to pain (congenital insensitivity to pain) or also including large fibre modalities such as touch (hereditary sensory neuropathy). Mutations in ion channels are increasingly recognised as causing functional disorders of somatosensation [123]. For instance homozygous loss of function mutations in *SCN9A*, the gene that encodes the sodium channel (Na_v) 1.7, have been shown to cause congenital insensitivity to pain [126]. Conversely, heterozygous gain of function variants are associated with a number of inherited pain disorders that include inherited erythromelalgia (IEM) [127] and paroxysmal extreme pain disorder (PEPD) [128]. *SCN9a* variants have also been linked to idiopathic small fibre neuropathy [129]. Some of these variants are relatively common in the general population and is likely to represent a gene environment interaction.

Our goals were to aid genetics diagnosis of patients with NPD within the UK, to determine the prevalence of known mutations associated with NPD in relation to distinct clinical presentations and finally to discover novel mutations causing NPD. The aim of our study was to determine whether genes previously implicated in neuropathic pain caused their clinical presentation. We recruited singleton individuals with extreme neuropathic pain phenotypes (both sensory loss and gain), all within the UK, from secondary care clinics located in Oxford, London, Salford, and Newcastle. We included participants older than 18 years of age with a proven history of life-style altering sensory disorder (either pain or loss of sensation) for greater than three months. The criteria for case definition for different clinical presentations are shown in **Supplementary Table 1 (NPD Criteria – Diagnostic Criteria)**. We excluded patients with a known underlying genetic cause of chronic pain, e.g. Fabry's disease and *SCN9A* congenital erythromelalgia although genetic pre-screening for these disorders was not mandatory. Patients with learning disorder or/and autistic features sufficient that they could not give either consent or partake in possible additional pain phenotyping were also excluded. The outcome measures used for patient phenotyping are shown in **Supplementary Table 1 (NPD Criteria Outcome Measures)**. The

Neuropathic Pain Special Interest Group (NeuPSIG) of the International Association for the Study of Pain (IASP)'s grading for neuropathic pain [6] was used to grade neuropathic pain for all study participants recruited.

A total of 193 study participants, with whole genome sequencing data available, underwent neuropathic pain grading. Excluded are five of the participants that were unaffected family members and one participant where phenotypic data was not available.

1. **No Neuropathic pain** – 8 (4.1%) participants did not report neuropathic pain
2. **Neuropathic pain unlikely** – 1 (0.5%) participant's history and pain distribution was not consistent with neuropathic pain.
3. **Possible Neuropathic pain** – 11 (5.7%) participants reported an appropriate history of a relevant lesion or disease, AND pain with a distinct neuroanatomically plausible distribution.
4. **Probable Neuropathic pain** – 56 (29.0%) participants satisfied criteria for possible neuropathic pain AND had clinical signs in the neuroanatomical distribution of neuropathic pain.
5. **Definite Neuropathic pain** – 111 (57.5%) participants satisfied criteria for probable neuropathic pain AND a diagnostic test confirmed a lesion of the somatosensory nervous system.

Appendix 2: Extreme red cell traits in UK Biobank

The UK Biobank is a biomedical cohort of approximately half a million participants, recruited in the UK between 2006 and 2010 [1]. The participants, 54% of whom are women, were aged between 37 and 73 years at their date of recruitment. Each participant underwent a baseline assessment at one of 21 centres across Great Britain, during which 4 ml of EDTA treated peripheral whole blood was collected for FBC analysis [12]. These blood samples were stored at 4 degrees centigrade and transported overnight in temperature controlled shipping boxes to the UK Biocentre laboratory in Stockport, Greater Manchester, UK, where FBCs were measured using a bank of four Beckman Coulter LH-700 instruments.

A GWAS of the FBC traits based on approximately 173,000 of the European ancestry participants in the UK Biobank cohort (n~133,000) and the INTERVAL trial (n~40,000) [90] has previously identified 582 genetic variants independently associated with quantitative properties of mature red blood cells [91]. Most of these variants were common, with only 40 having an in sample MAF lower than 1%. The GWAS design had limited power to identify associations with rare variants for three reasons. Firstly, it relied on the imputation of rare alleles from the UK10K/1000 Genomes reference panels [44, 26], which are too small to contain a large proportion of the rare haplotypes carried by the hundreds of thousands of GWAS participants. Secondly, in an attempt to inhibit spurious associations due to statistical model misspecification, participants with extreme phenotype data were deliberately excluded from the association analyses. A genetic association with a rare variant can only be detected with high probability if its effect size is large, which implies that carriers of rare alleles exhibiting detectable associations were more likely than typical study participants to have been excluded from the GWAS analyses. Thirdly, the GWAS relied on univariable genetic analyses to identify allelic associations, and these can have less power than methods such as BeviMed, which are able to model jointly the association of multiple rare variants in a DNA sequence element [89].

We sought to identify a subset of UK Biobank participants likely to carry rare alleles modulating properties of peripheral red blood cells, which could plausibly be identified by whole genome sequencing. Our method was to construct a univariable score from the UK Biobank baseline mature red cell FBC measurements, which we thought likely to have high rare-variant heritability. We then selected participants for sequencing from each tail of the distribution of the score.

To construct the score, we used the 65 variants with MAF < 1% that were reported to be significantly ($P < 8.31 \times 10^{-9}$) associated with at least one of twelve quantitative properties of (mature or immature) red cells by [91], as a model for the likely effect of rare alleles on the baseline UK Biobank FBC. **Figure 4a** shows the pairwise relationships between the estimated effect sizes of these variants on the red cell FBC parameters MCV, RBC#, HGB and RDW. This subset of parameters is minimal in the sense that the other mature red blood cell FBC parameters can be calculated deterministically from them. The estimated effect sizes were reported by [91] as per allele additive differences in the mean of the rank-inverse standard unit normalised trait and are therefore given in units which are comparable across traits. In general, MCV and RBC# exhibit a greater range and variance in absolute effect size than HGB and RDW. 91 reported that, of all the traits they studied, MCV has the highest estimated common variant heritability and that it yielded the second largest number of associated variants with MAF < 1%. It seems reasonable to conjecture from this, that MCV also has a relatively high *rare* variant heritability. There is a strong inverse correlation between the effect sizes of alleles perturbing MCV and RBC#, suggesting that the effect sizes may measure aspects of the same underlying biological mechanism, perhaps the control of the total blood volume proportion of red cells (haematocrit). Since we could not identify any other precise systematic relationship between aspects of the joint distribution of the four estimated red cell trait effect sizes, we decided to restrict our attention to the marginal joint distribution of MCV and RBC# effect sizes (highlighted by the red square, **Figure 4a**). We used Deming regression to estimate the approximate linear relationship:

$$\beta MCV \quad = -1.69 \times \beta RBC\# \qquad (1)$$

between the effect sizes for the two traits. This linear relationship is shown by a red line in **Figure 4a.**

We took the baseline UK Biobank MCV and RBC# parameters and adjusted them to remove the effect of various sources of technical and biological variation. A detailed description of the adjustments can be found in the the STAR methods section of [91]. In brief, the phenotypes were firstly adjusted to remove differences between instruments, to remove time dependent instrument drifts and to remove the effect of delay time between venepuncture and measurement. Data acquired on days where the instrument mean was an outlier for the corresponding trait were removed. Subsequently, participants who had a hysterectomy or who had a self-report or medical history containing a record of myelofibrosis, lymphoma, leukemia, malignant lymphoma, multiple myeloma, multiple myelofibrosis or myelodysplasia, chronic lymphocytic leukemia, chronic myeloid leukemia, acute myeloid leukemia, polycythemia vera, polycythemia, a myeloproliferative disorder, essential thrombocytosis, a haematological cancer histology report, an unspecified lymphatic or general haematological neoplasm, a myelodysplastic syndrome, or an unspecified heme malignancy, monoclonal gammopathy, an unspecified hereditary haematological disorder, haemochromatosis, thalassaemia, haemophilia, sickle cell anaemia, neutropenia, lymphopenia or pancytopenia were excluded from analysis. Finally, the phenotypes were adjusted in a second stage to remove the effects of sex, age, menopause status, the interaction of sex and menopause status with age, height, weight and for the effects of history and current habits of smoking and alcohol consumption.

We excluded all participants who did not self report their ancestry as one of “British”, “White”, “Irish” or “Any other white background” in the UK Biobank baseline assessment questionnaire. We also excluded individuals whose genotypes appeared in the interim 2015 genetic data release and who were identified as having non-European ancestry by the principal components approach reported in [91].

We defined the selection score S_i for the *i*th UK Biobank participant as the (signed) Euclidean distance in $\Re^2$ between the origin and the point $P(f_i,g_i)$, where P is the orthogonal projection onto the line*:*

$$g = -1.69 \times f, \qquad (2)$$

where $f_i = f(RBC\#_i)$ and $g_i = g(MCV_i)$ for functions f and g that Box-Cox transform, standardise and centre the technically and biologically adjusted traits *RBC#* and *MCV* respectively, in the UK Biobank participants not hitherto excluded. In [91] the covariate adjusted traits were rank inverse normalised against $N(0,1)$ before the genetic association analyses. However, here we preferred to work with Box-Cox transformed traits in order to adjust the central part of the data towards a Gaussian without over shrinking outliers, which might reduce power. The selection score can be expressed as:

$$S_i \equiv -1.69 \times g(MCV_i) + f(RBC\#_i)$$

and its, centered and standardised distribution in male and post menopausal female UK Biobank participants is shown in sub-panel bounded by a dotted line in **Figure 4b.**

We excluded pre-menopausal females, as candidates for sequencing because of their high prevalence of anemia and because of the additional component of non-genetic variation in each red cell parameter that is induced by the menstrual cycle. We also excluded individuals with a UK Biobank report of an insufficient DNA quantity (less than 4.7ug) to generate a working stock of 130ul of 36ng/ul(TRINEAN measured concentration < 36ng/μl, PicoGreen measured concentration < 36ng/μl) or with a UK Biobank report of inconsistency between genetic and self reported sex. Finally, we excluded participants with a platelet count below 75x10^9/l, a white blood cell count below 0.5 x10^9/l or a white blood cell count above 13x10^9/l.

We partitioned the 316,739 male and post-menopausal female participants with a computed score value into six gender specific age groups thresholding at 53.3 years and 62 years. (These age groups divide those participants with a score, including the pre-menopausal women, into three groups, each of approximately 125,000 participants). Within each of these six sex-age groups, we ranked the study participants according to the value of the score. We selected a total of 384 individuals from the tail of each score, stratifying the selection by sex-age group so that the final selection for each tail sampled each group in proportion to its size. This additional stratification was necessary despite the adjustment of the mean of each trait for age and sex to ensure reasonable age and sex balance in the tails. The full blood count of each selected participant was reviewed by an expert panel of haematologists to exclude any text-book non-genetic or somatic pathologies such as bone-marrow failure, polycythemia vera or essential thrombocytopenia, which might explain the extreme value of the selection score. A small sample of DNA was screened by the Cambridge Blood and Stem-Cell Biobank for the JAK2 mutation V617F, a common cause of somatic myeloproliferative disorders. Any participants failing the FBC or DNA screen were replaced by the next most extreme individual in the same sex-age subgroup.

DNA samples from a total of 416 male and 352 female UK Biobank participants were retrieved from the central sample archive and sent to Illumina sequencing. The main panel of **Figure 4b** shows the distribution of the score of the selected individuals for each tail, while **Figure 4c** is a bivariate scatter showing the distribution of RBC# and MCV (after adjustment for technical but not biological variation) in the two tails. Of the 768 individuals sent for whole genome sequencing one individual from each tail failed Illumina sequencing quality control and two distinct individuals from the right tail failed in-house checks for DNA contamination.

References

1. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015 Mar;12(3):e1001779

2. Westbury SK, Turro E, Greene D, Lentaigne C, Kelly AM, Bariana TK, et al. Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. Genome Med. 2015;7(1):36

3. Beighton P, De Paepe A, Steinmann B, Tsipouras P, Wenstrup RJ. Ehlers-Danlos syndromes: revised nosology, Villefranche, 1997. Ehlers-Danlos National Foundation (USA) and Ehlers-Danlos Support Group (UK). Am J Med Genet. 1998 Apr 28;77(1):31-7

4. Carss KJ, Arno G, Erwood M, Stephens J, Sanchis-Juan A, Hull S, et al. Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. Am J Hum Genet. 2017 Jan 5;100(1):75-90

5. International rules for multiple primary cancers (ICD-0 third edition). Eur J Cancer Prev. 2005 Aug;14(4):307-8

6. Finnerup NB, Haroutounian S, Kamerman P, Baron R, Bennett DL, Bouhassira D, et al. Neuropathic pain: an updated grading system for research and clinical practice. Pain. 2016 Aug;157(8):1599-606

7. Hadinnapola C, Bleda M, Haimel M, Screaton N, Swift A, Dorfmuller P, et al. Phenotypic Characterization of EIF2AK4 Mutation Carriers in a Large Cohort of Patients Diagnosed Clinically With Pulmonary Arterial Hypertension. Circulation. 2017 Nov 21;136(21):2022-2033

8. Gräf S, Haimel M, Bleda M, Hadinnapola C, Southgate L, Li W, et al. Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. Nat Commun. 2018 Apr 12;9(1):1416

9. Rhodes CJ, Batai K, Bleda M, Haimel M, Southgate L, Germain M, et al. Genetic determinants of risk in pulmonary arterial hypertension: international genome-wide association studies and meta-analysis. Lancet Respir Med. 2018 Dec 5;

10. Tuijnenburg P, Lango Allen H, Burns SO, Greene D, Jansen MH, Staples E, et al. Loss-of-function nuclear factor kappaB subunit 1 (NFKB1) variants are the most common monogenic cause of common variable immunodeficiency in Europeans. J Allergy Clin Immunol. 2018 Oct;142(4):1285-1296

11. Goodship TH, Cook HT, Fakhouri F, Fervenza FC, Fremeaux-Bacchi V, Kavanagh D, et al. Atypical hemolytic uremic syndrome and C3 glomerulopathy: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int. 2017 Mar;91(3):539-551

12. Elliott P, Peakman TC. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. Int J Epidemiol. 2008 Apr;37(2):234-44

13. Office for National Statistics, 2011 Census, https://www.ons.gov.uk/census/2011census, 2011.

14. Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet. 2008 Nov;83(5):610-5

15. Haendel MA, Chute CG, Robinson PN. Classification, Ontology, and Precision Medicine. N Engl J Med. 2018 Oct 11;379(15):1452-1462

16. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics. 2013 Aug 15;29(16):2041-3

17. Genome Reference Consortium, GRCh37, https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/, 2009.

18. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9

19. Illumina, Starling (Isaac Variant Caller), https://support.illumina.com/help/BS_App_TS_Amplicon_OLH_15055858/Content/Source/Informatics/Apps/IsaacVariantCaller_appENR.htm, 2016.

20. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156-8

21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559-75

22. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285-91

23. Gudbjartsson DF, Sulem P, Helgason H, Gylfason A, Gudjonsson SA, Zink F, et al. Sequence variants from whole genome sequencing a large group of Icelanders. Sci Data. 2015;2:150011

24. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet. 2012 Nov 2;91(5):839-48

25. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203-209

26. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68-74

27. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010 Nov 15;26(22):2867-73

28. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012 Dec 15;28(24):3326-8

29. Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free Estimation of Recent Genetic Relatedness. Am J Hum Genet. 2016 Jan 7;98(1):127-48

30. Staples J, Nickerson DA, Below JE. Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. Genet Epidemiol. 2013 Feb;37(2):136-41

31. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987-93

32. Google, Protocol Buffers, https://developers.google.com/protocol-buffers, 2018.

33. Medina, I. et al., OpenCGA, http://docs.opencb.org/display/opencga, 2017.

34. The Apache Software Foundation, Apache HBase, https://hbase.apache.org/, 2018.

35. Handsaker, B., HTSJDK, https://raw.githubusercontent.com/samtools/htsjdk/master/src/main/java/htsjdk/tribble/util/popgen/HardyWeinbergCalculation.java, 2009.

36. Bleda M, Tarraga J, de Maria A, Salavert F, Garcia-Alonso L, Celma M, et al. CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W609-14

37. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar;46(3):310-5

38. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812-4

39. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan;Chapter 7:Unit7.20

40. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005 Jul;15(7):901-13

41. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005 Aug;15(8):1034-50

42. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010 Jan;20(1):110-21

43. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015 Oct 1;526(7571):75-81

44. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015 Oct 1;526(7571):82-90

45. Song W, Gardner SA, Hovhannisyan H, Natalizio A, Weymouth KS, Chen W, et al. Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification. Genet Med. 2016 Aug;18(8):850-4

46. NHLBI, Trans-Omics for Precision Medicine (TOPMed) program, https://www.nhlbiwgs.org/, 2018.

47. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017 Jun;136(6):665-677

48. Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I., Spark: Cluster Computing with Working Sets, https://www.usenix.org/legacy/event/hotcloud10/tech/full_papers/Zaharia.pdf, 2010.

49. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014 Apr 1;30(7):1006-7

50. EMBL-EBI Ensembl FAQ, What do the different biotypes in Ensembl mean?, https://www.ensembl.org/Help/Faq?id=468, 2018.

51. EMBL-EBI Ensembl, Ensembl Variation - Calculated variant consequences, http://www.ensembl.org/info/genome/variation/prediction/predicted_data.html, 2018.

52. Taylor MS, Ponting CP, Copley RR. Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. Genome Res. 2004 Apr;14(4):555-66

53. Lopez J, Coll J, Haimel M, Kandasamy S, Tarraga J, Furio-Tari P, et al. HGVA: the Human Genome Variation Archive. Nucleic Acids Res. 2017 Jul 3;45(W1):W189-W194

54. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016 Apr 15;32(8):1220-2

55. Roller E, Ivakhno S, Lee S, Royce T, Tanner S. Canvas: versatile and scalable detection of copy number variants. Bioinformatics. 2016 Aug 1;32(15):2375-7

56. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007 Nov;17(11):1665-74

57. University of Washington Genome Sciences, Segmental Duplication DB, http://humanparalogy.gs.washington.edu/build37/build37.htm, 2009.

58. She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, et al. Shotgun sequence assembly and recent segmental duplications within the human genome. Nature. 2004 Oct 21;431(7011):927-30
59. MacArthur JA, Morales J, Tully RE, Astashyn A, Gil L, Bruford EA, et al. Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. Nucleic Acids Res. 2014 Jan;42(Database issue):D873-8
60. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016 Jun 6;17(1):122
61. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012 Jul 6;337(6090):64-9
62. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016 Jan 4;44(D1):D862-8
63. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015 Jan;43(Database issue):D789-98
64. National Center for Biotechnology Information, U.S. National Library of Medicine, PubMed, https://www.ncbi.nlm.nih.gov/pubmed, 2018.
65. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015 May;17(5):405-24
66. Sanchis-Juan A, Stephens J, French CE, Gleadall N, Megy K, Penkett C, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. Genome Med. 2018 Dec 7;10(1):95
67. Tartaglia M, Mehler EL, Goldberg R, Zampino G, Brunner HG, Kremer H, et al. Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. Nat Genet. 2001 Dec;29(4):465-8
68. Tartaglia M, Kalidas K, Shaw A, Song X, Musat DL, van der Burgt I, et al. PTPN11 mutations in Noonan syndrome: molecular spectrum, genotype-phenotype correlation, and phenotypic heterogeneity. Am J Hum Genet. 2002 Jun;70(6):1555-63
69. Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, Guo H, et al. An evaluation of copy number variation detection tools from whole-exome sequencing data. Hum Mutat. 2014 Jul;35(7):899-907
70. Pasmant E, Goussard P, Baranes L, Laurendeau I, Quentin S, Ponsot P, et al. First description of ABCB4 gene deletions in familial low phospholipid-associated cholelithiasis and oral contraceptives-induced cholestasis. Eur J Hum Genet. 2012 Mar;20(3):277-82
71. Anzivino C, Odoardi MR, Meschiari E, Baldelli E, Facchinetti F, Neri I, et al. ABCB4 and ABCB11 mutations in intrahepatic cholestasis of pregnancy in an Italian population. Dig Liver Dis. 2013 Mar;45(3):226-32
72. Larkin EK, Newman JH, Austin ED, Hemnes AR, Wheeler L, Robbins IM, et al. Longitudinal analysis casts doubt on the presence of genetic anticipation in heritable pulmonary arterial hypertension. Am J Respir Crit Care Med. 2012 Nov 1;186(9):892-6

73. Koziell A, Grech V, Hussain S, Lee G, Lenkkeri U, Tryggvason K, et al. Genotype/phenotype correlations of NPHS1 and NPHS2 mutations in nephrotic syndrome advocate a functional inter-relationship in glomerular filtration. Hum Mol Genet. 2002 Feb 15;11(4):379-88
74. Tory K, Menyhard DK, Woerner S, Nevo F, Gribouval O, Kerti A, et al. Mutation-dependent recessive inheritance of NPHS2-associated steroid-resistant nephrotic syndrome. Nat Genet. 2014 Mar;46(3):299-304
75. Hjorten R, Skorecki K. Leveraging Ancestral Heterogeneity to Map Shared Genetic Risk Loci in Pediatric Steroid-Sensitive Nephrotic Syndrome. J Am Soc Nephrol. 2018 Jul;29(7):1793-1794
76. Meyer E, Carss KJ, Rankin J, Nichols JM, Grozeva D, Joseph AP, et al. Mutations in the histone methyltransferase gene KMT2B cause complex early-onset dystonia. Nat Genet. 2017 Feb;49(2):223-237
77. Stritt S, Nurden P, Turro E, Greene D, Jansen SB, Westbury SK, et al. A gain-of-function variant in DIAPH1 causes dominant macrothrombocytopenia and hearing loss. Blood. 2016 Jun 9;127(23):2903-14
78. Westbury SK, Downes K, Burney C, Lozano ML, Obaji SG, Toh CH, et al. Phenotype description and response to thrombopoietin receptor agonist in DIAPH1-related disorder. Blood Adv. 2018 Sep 25;2(18):2341-2346
79. Turro E, Greene D, Wijgaerts A, Thys C, Lentaigne C, Bariana TK, et al. A dominant gain-of-function mutation in universal tyrosine kinase SRC causes thrombocytopenia, myelofibrosis, bleeding, and bone pathologies. Sci Transl Med. 2016 Mar 2;8(328):328ra30
80. Noetzli L, Lo RW, Lee-Sherick AB, Callaghan M, Noris P, Savoia A, et al. Germline mutations in ETV6 are associated with thrombocytopenia, red cell macrocytosis and predisposition to lymphoblastic leukemia. Nat Genet. 2015 May;47(5):535-538
81. Zhang MY, Churpek JE, Keel SB, Walsh T, Lee MK, Loeb KR, et al. Germline ETV6 mutations in familial thrombocytopenia and hematologic malignancy. Nat Genet. 2015 Feb;47(2):180-5
82. Song WJ, Sullivan MG, Legare RD, Hutchings S, Tan X, Kufrin D, et al. Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. Nat Genet. 1999 Oct;23(2):166-75
83. Poggi M, Canault M, Favier M, Turro E, Saultier P, Ghalloussi D, et al. Germline variants in ETV6 underlie reduced platelet formation, platelet dysfunction and increased levels of circulating CD34+ progenitors. Haematologica. 2017 Feb;102(2):282-294
84. Evans JD, Girerd B, Montani D, Wang XJ, Galie N, Austin ED, et al. BMPR2 mutations and survival in pulmonary arterial hypertension: an individual participant data meta-analysis. Lancet Respir Med. 2016 Feb;4(2):129-37
85. Hancks DC, Kazazian HH Jr. Roles for retrotransposon insertions in human disease. Mob DNA. 2016;7:9
86. Cartron JP. Rh-deficiency syndrome. Lancet. 2001 Dec;358 Suppl:S57
87. Mouro-Chanteloup I, D'Ambrosio AM, Gane P, Le Van Kim C, Raynal V, Dhermy D, et al. Cell-surface expression of RhD blood group polypeptide is posttranscriptionally regulated by the RhAG glycoprotein. Blood. 2002 Aug 1;100(3):1038-47

88. Kulkarni SS, Vasantha K, Gogri H, Parchure D, Madkaikar M, Ferec C, et al. First report of Rhnull individuals in the Indian population and characterization of the underlying molecular mechanisms. Transfusion. 2017 Aug;57(8):1944-1948
89. Greene D, Richardson S, Turro E. A Fast Association Test for Identifying Pathogenic Variants Involved in Rare Diseases. Am J Hum Genet. 2017 Jul 6;101(1):104-114
90. Moore C, Sambrook J, Walker M, Tolkien Z, Kaptoge S, Allen D, et al. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. Trials. 2014 Sep 17;15:363
91. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016 Nov 17;167(5):1415-1429.e19
92. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat. 2015 Oct;36(10):928-30
93. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. Hum Mutat. 2015 Oct;36(10):915-21
94. Jones CI, Bray S, Garner SF, Stephens J, de Bono B, Angenent WG, et al. A functional genomics approach reveals novel quantitative trait loci associated with platelet signaling pathways. Blood. 2009 Aug 13;114(7):1405-16
95. Ito Y, Carss KJ, Duarte ST, Hartley T, Keren B, Kurian MA, et al. De Novo Truncating Mutations in WASF1 Cause Intellectual Disability with Seizures. Am J Hum Genet. 2018 Jul 5;103(1):144-153
96. Dahl JP, Wang-Dunlop J, Gonzales C, Goad ME, Mark RJ, Kwak SP. Characterization of the WAVE1 knock-out mouse: implications for CNS development. J Neurosci. 2003 Apr 15;23(8):3343-52
97. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016 Nov 17;167(5):1369-1384.e19
98. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60
99. Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics. 2008 Nov 1;24(21):2537-8
100. Blueprint Epigenome, ChIP-Seq Analysis Pipeline, http://dcc.blueprint-epigenome.eu/#/md/chip_seq_grch37, 2016.
101. The Broad Institute, Picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF., http://broadinstitute.github.io/picard, 2018.
102. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016 Jul 8;44(W1):W160-5
103. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137

104. Skultetyova L, Ustinova K, Kutil Z, Novakova Z, Pavlicek J, Mikesova J, et al. Human histone deacetylase 6 shows strong preference for tubulin dimers over assembled microtubules. Sci Rep. 2017 Sep 14;7(1):11547

105. Petersen R, Lambourne JJ, Javierre BM, Grassi L, Kreuzhuber R, Ruklisa D, et al. Platelet function is modified by common sequence variation in megakaryocyte super enhancers. Nat Commun. 2017 Jul 13;8:16058

106. Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, et al. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science. 2016 Nov 11;354(6313):769-773

107. Freson K, Devriendt K, Matthijs G, Van Hoof A, De Vos R, Thys C, et al. Platelet characteristics in patients with X-linked macrothrombocytopenia because of a novel GATA1 mutation. Blood. 2001 Jul 1;98(1):85-92

108. Wijgaerts A, Wittevrongel C, Thys C, Devos T, Peerlinck K, Tijssen MR, et al. The transcription factor GATA1 regulates NBEAL2 expression through a long-distance enhancer. Haematologica. 2017 Apr;102(4):695-706

109. Zhang Y, Kwon S, Yamaguchi T, Cubizolles F, Rousseaux S, Kneissel M, et al. Mice lacking histone deacetylase 6 have hyperacetylated tubulin but are viable and develop normally. Mol Cell Biol. 2008 Mar;28(5):1688-701

110. Fukada M, Hanai A, Nakayama A, Suzuki T, Miyata N, Rodriguiz RM, et al. Loss of deacetylation activity of Hdac6 affects emotional behavior in mice. PLoS One. 2012;7(2):e30924

111. Sadoul K, Wang J, Diagouraga B, Vitte AL, Buchou T, Rossini T, et al. HDAC6 controls the kinetics of platelet activation. Blood. 2012 Nov 15;120(20):4215-8

112. Messaoudi K, Ali A, Ishaq R, Palazzo A, Sliwa D, Bluteau O, et al. Critical role of the HDAC6-cortactin axis in human megakaryocyte maturation leading to a proplatelet-formation defect. Nat Commun. 2017 Nov 27;8(1):1786

113. de Waele L, Freson K, Louwette S, Thys C, Wittevrongel C, de Vos R, et al. Severe gastrointestinal bleeding and thrombocytopenia in a child with an anti-GATA1 autoantibody. Pediatr Res. 2010 Mar;67(3):314-9

114. Heremans J, Garcia-Perez JE, Turro E, Schlenner SM, Casteels I, Collin R, et al. Abnormal differentiation of B cells and megakaryocytes in patients with Roifman syndrome. J Allergy Clin Immunol. 2018 Aug;142(2):630-646

115. Di Michele M, Thys C, Waelkens E, Overbergh L, D'Hertog W, Mathieu C, et al. An integrated proteomics and genomics analysis to unravel a heterogeneous platelet secretion defect. J Proteomics. 2011 May 16;74(6):902-13

116. Lopez-Herrera G, Tampella G, Pan-Hammarstrom Q, Herholz P, Trujillo-Vargas CM, Phadwal K, et al. Deleterious mutations in LRBA are associated with a syndrome of immune deficiency and autoimmunity. Am J Hum Genet. 2012 Jun 8;90(6):986-1001

117. Wendling F, Maraskovsky E, Debili N, Florindo C, Teepe M, Titeux M, et al. cMpl ligand is a humoral regulator of megakaryocytopoiesis. Nature. 1994 Jun 16;369(6481):571-4

118. Tijssen MR, di Summa F, van den Oudenrijn S, Zwaginga JJ, van der Schoot CE, Voermans C, et al. Functional analysis of single amino-acid mutations in the thrombopoietin-receptor Mpl underlying congenital amegakaryocytic thrombocytopenia. Br J Haematol. 2008 Jun;141(6):808-13

119. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, et al. MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics. 2005 Jul 1;21(13):2933-42
120. Genalice, Modular high performance NGS secondary analysis suite, http://www.genalice.com/solutions/products/, 2018.
121. Pluss M, Kopps AM, Keller I, Meienberg J, Caspar SM, Dubacher N, et al. Need for speed in accurate whole-genome data analysis: GENALICE MAP challenges BWA/GATK more than PEMapper/PECaller and Isaac. Proc Natl Acad Sci U S A. 2017 Oct 3;114(40):E8320-E8322
122. Treede RD, Jensen TS, Campbell JN, Cruccu G, Dostrovsky JO, Griffin JW, et al. Neuropathic pain: redefinition and a grading system for clinical and research purposes. Neurology. 2008 Apr 29;70(18):1630-5
123. Bennett DL, Woods CG. Painful and painless channelopathies. Lancet Neurol. 2014 Jun;13(6):587-99
124. Geha P, Yang Y, Estacion M, Schulman BR, Tokuno H, Apkarian AV, et al. Pharmacotherapy for Pain in a Family With Inherited Erythromelalgia Guided by Genomic Analysis and Functional Profiling. JAMA Neurol. 2016 Jun 1;73(6):659-67
125. Blesneac I, Themistocleous AC, Fratter C, Conrad LJ, Ramirez JD, Cox JJ, et al. Rare NaV1.7 variants associated with painful diabetic peripheral neuropathy. Pain. 2018 Mar;159(3):469-480
126. Cox JJ, Reimann F, Nicholas AK, Thornton G, Roberts E, Springell K, et al. An SCN9A channelopathy causes congenital inability to experience pain. Nature. 2006 Dec 14;444(7121):894-8
127. Yang Y, Wang Y, Li S, Xu Z, Li H, Ma L, et al. Mutations in SCN9A, encoding a sodium channel alpha subunit, in patients with primary erythermalgia. J Med Genet. 2004 Mar;41(3):171-4
128. Fertleman CR, Baker MD, Parker KA, Moffatt S, Elmslie FV, Abrahamsen B, et al. SCN9A mutations in paroxysmal extreme pain disorder: allelic variants underlie distinct channel defects and phenotypes. Neuron. 2006 Dec 7;52(5):767-74
129. Faber CG, Hoeijmakers JG, Ahn HS, Cheng X, Han C, Choi JS, et al. Gain of function Nanu1.7 mutations in idiopathic small fiber neuropathy. Ann Neurol. 2012 Jan;71(1):26-39

Affiliation numbering continues from the main paper

NIHR BioResource - Rare Diseases Collaborators: Zoe Adhya[168], Maryam Afzal[43], Irshad Ahmed[169], Saeed Ahmed[227], Jayanthi Alamelu[45], Raza Alikhan[104], Louise Allen[9,83,228], Arif Alvi[108], Gautam Ambegaonkar[229], Ariharan Anantharachagan[36,230], Gururaj Arumugakani[231], Rita Arya[232], Steve Austin[45], Yesim Aydinok[233], Waqar Ayub[234], Mohsin Badat[34], Trevor Baglin[36], Jonathan Barratt[235], John Baski[118,119], Rachel Bates[34], Gareth Baynam[236,237,238], Claire Bethune[239], Neha Bhatnagar[107], Shahnaz Bibi[47], Preetham Boddana[240], Claire Booth[47], Angela Brady[241], Annette Briley[17], Richard Brown[242], Christine Bryson[1,2], Jackie Buck[243], Gary Campbell[244], Natalie Canham[141,241], Jenny Carmichael[36], Elizabeth Chalmers[170], Melissa V Chan[245], Anita Chandra[36], Sam Chong[153], Emma M Clement[47], Virginia Clowes[241], Victoria Cookson[47], Amanda Creaser-Myers[246], Rosa Da Costa[119], Sophie Davies[36], Sarah Deacock[247],

Patrick B Deegan[27], John Dempster[168], Michael Desborough[34], Lisa A Devlin[163], Anand Dixit[117], Rainer Doffinger[158], Helen Dolling[1,2], Natalie Dormand[119], Tariq El-Shanawany[172], Tony Elston[248], Ingrid Emmerson[117], Henry Farmery[6], Helen Firth[3,49], Nick Fordham[34], Bruce Furie[105], Alice Gardham[241], H Bobby Gaspar[31], Johanna Gebhart[249], Neeti Ghali[250], Rohit Ghurye[168], Rodney D Gilbert[251,252], Lionel Ginsberg[55,106,161], Joanna C Girling[253], Paul Gissen[47,55], Kathleen M Gorman[147,148], Alan Greenhalgh[254], Sian Griffiths[255], Yisu Gu[34], Robert D M Hadden[256], Csaba Halmagyi[1,2], Tracey Hammerton[1,2], Lorraine Harper[157], Claire Harrison[122], Shivaram Hegde[255], Robert H Henderson[47], Anke Hensiek[36], Yvonne M C Henskens[257], Muriel Holder[122], Sean Hughes[230], Stephen Hughes[258], Anna E Huis in 't Veld[188], Jane A Hurst[47], Val Irvine[189], Praveen Jeevaratnam[259], Mark Johnson[260], Bryony Jones[261], Caroline Jones[262], Yousuf Karim[181,247], Mahantesh Karoshi[263], David Keeling[107], Fiona Kennedy[189], Sorena Kiani[168], Andrew King[34], Sally Kinsey[264], Alison Kirkpatrick[181], Nigel Klein[47], Ellen Knox[265], Deepa Krishnakumar[36], James Laffan[168], Sarah H A Lawman[266], Sara E Lear[36,244,267], Melissa Lees[47], Andrew Lewington[268], James Liang[269], Ri Liesner[102], Malcolm Macdougall[117], Rajiv D Machado[270,271], Lucy H Mackillop[34,272], Robert MacLaren[50], Laura Magee[273], Mohamed Mahdi-Rogers[121], Mike Makris[202,246], Ania Manson[36], Adnan Manzur[47], Patrick B Mark[179,274], Larahmie Masati[66], Vera Matser[1,2], Anna Maw[36], Elizabeth M McDermott[162], Simon J McGowan[28,30], Coleen McJannet[1,2], Amy McTague[147,148], Sharon Meehan[66], Catherine L Mercer[88], Anna C Michell[32,34], David Milford[275], Anoop Mistry[231], Jason Moore[276], Valerie Morrisson[36], Sai H K Murng[178,179], Elaine Murphy[106], Joanne Ng[147,148], Adeline Ngoh[147,148], Muna Noori[261], Eric Oksenhendler[277], Albert C M Ong[165,202], Shokri Othman[66], Antonis Pantazis[119], Apostolos Papandreou[147,148,278], Alasdair P J Parker[36], Georgina Parsons[34], K John Pasi[279], Chris Patch[122], Jeanette H Payne[280], David Perry[69], Bartlomiej Piechowski-Jozwiak[121], Fernando Pinto[170], Gary J Polwarth[159], Mark J Ponsford[281,282], Sanjay Prasad[118,119], Waseem Qasim[31,47], Ellen Quinn[47], Isabella Quinti[283], Sanjay Raina[284], Lavanya Ranganathan[66], Julia Rankin[276], Karola Rehnstrom[1,2], Evan Reid[9,56], Mary M Reilly[55,153], Shoshana Revel-Vilk[285], Emma E Richards[111], Mike Richards[286], Matthew T Rondina[287], Elisabeth Rosser[47], Peter Rothwell[288], Richard Sandford[56], Saikat Santra[275], Gwen Schotte[188], Harald Schulze[289], Suranjith L Seneviratne[52,161], Fiona Shackley[165], Pankaj Sharma[290], Hassan Shehata[171,291], Deborah Shipley[254], Manish D Sinha[19,20,122], Linda Sneddon[116], Aman Sohal[275], Laura Southgate[271,292], Miranda Splitt[117], Hans Stauss[161], Cathal L Steele[293], Penelope E Stein[121], Sophie Stock[1,2], Matthew J Stubbs[23,24], Emily Symington[69], Gordon B Taylor[294], Jecko Thachil[295], Dorothy A Thompson[47], Sarah Trippier[273], Rafal Urniaz[27], Marijcke W M Veltman[1,2], Julie Vogt[139], Ajay Vora[296], Minka J A Vries[257], Emma L Wakeling[241], Roddy Walsh[118,119], Ivy Wanjiku[66], Timothy Warner[168], Evangeline Wassmer[275], Henry G Watson[294], Dean Waugh[113], Nick Webb[258], Angela Welch[112], David Werring[106], Lisa Willcocks[36], David J Williams[106], Henna Wong[34], Sarita Workman[161], Nigel Yeatman[168]
[1]Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [2]NIHR BioResource, Cambridge University Hospitals NHS Foundation, Cambridge Biomedical Campus, Cambridge, UK. [3]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [6]MRC Biostatistics Unit, Cambridge Institute of Public Health, University of Cambridge, Cambridge, UK. [9]Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [17]Women and Children's Health, School of Life Course Sciences, King's College London, London, UK. [19]King's College London, London, UK. [20]Department of

Paediatric Nephrology, Evelina London Children's Hospital, Guy's & St Thomas' NHS Foundation Trust, London, UK. [23]Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [24]Centre for Haematology, Imperial College London, London, UK. [27]Department of Medicine, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [28]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [30]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals Trust, Oxford, UK. [31]UCL Great Ormond Street Institute of Child Health, London, UK. [32]Department of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [34]Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [36]Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [43]Bristol Renal and Children's Renal Unit, Bristol Medical School, University of Bristol, Bristol, UK. [45]Department of Haematology, Guy's and St Thomas' NHS Foundation Trust, London, UK. [47]Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [49]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [50]Moorfields Eye Hospital NHS Foundation Trust, London, UK. [52]Institute of Immunity and Transplantation, University College London, London, UK. [55]University College London, London, UK. [56]Department of Clinical Genetics, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [66]Department of Medicine, Imperial College London, London, UK. [69]Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [83]Department of Renal Medicine, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [88]Southampton General Hospital, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [102]Department of Haematology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [104]The Arthur Bloom Haemophilia Centre, University Hospital of Wales, Cardiff, UK. [105]Beth Israel Deaconess Medical Centre and Harvard Medical School, Boston, USA. [106]University College London Hospitals NHS Foundation Trust, London, UK. [107]Oxford Haemophilia and Thrombosis Centre, The Churchill Hospital, Oxford University Hospitals NHS Trust, Oxford, UK. [108]Glasgow Royal Infirmary, NHS Greater Glasgow and Clyde, Glasgow, UK. [111]Department of Neurology, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [112]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK. [113]Department of Neurology, Leeds Teaching Hospital NHS Trust, Leeds, UK. [116]Newcastle University, Newcastle upon Tyne, UK. [117]Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [118]National Heart and Lung Institute, Imperial College London, London, UK. [119]Royal Brompton Hospital, Royal Brompton and Harefield NHS Foundation Trust, London, UK. [121]King's College Hospital NHS Foundation Trust, London, UK. [122]Guy's and St Thomas' Hospital, Guy's and St Thomas' NHS Foundation Trust, London, UK. [139]West Midlands Regional Genetics Service, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK. [141]Department of Clinical Genetics, Liverpool Women's NHS Foundation, Liverpool, UK. [147]Developmental Neurosciences, UCL Great Ormond Street Institute of Child Health, London, UK. [148]Department of Neurology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [153]The National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK. [157]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [158]Division of Clinical Biochemistry and Immunology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [159]Royal Papworth Hospital NHS Foundation Trust, Cambridge, UK. [161]Royal Free London NHS Foundation Trust, London, UK. [162]Nottingham University Hospitals NHS Trust, Nottingham, UK. [163]Regional Immunology Service, The Royal Hospitals, Belfast, UK. [165]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [168]Barts Health NHS Foundation Trust, London, UK. [169]Birmingham Heartlands Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [170]Royal Hospital for Children, NHS Greater Glasgow and Clyde, Glasgow, UK. [171]Epsom & St Helier University Hospitals NHS Trust, London, UK. [172]Immunodeficiency Centre for Wales, University Hospital of Wales, Cardiff, IUK. [178]Gartnavel General Hospital, NHS Greater Glasgow and Clyde, Glasgow, UK. [179]Queen Elizabeth University Hospital, Glasgow, UK. [181]Frimley Park Hospital, NHS Frimley Health Foundation Trust, Camberley, UK. [188]Department of Pulmonary Medicine, VU University Medical Centre, Amsterdam, The Netherlands. [189]Golden Jubilee National Hospital, Glasgow, UK. [202]Department of Infection, Immunity & Cardiovascular Disease, University of Sheffield, Sheffield, UK. [227]Department of Renal Medicine, Sunderland Royal Hospital, Sunderland, UK. [228]Department of Ophthalmology, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [229]Child Development Centre, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [230]Lancashire Teaching Hospital NHS Foundation Trust, Lancashire, UK. [231]The Leeds Teaching Hospitals NHS Trust, Leeds, UK. [232]Warrington and Halton Hospitals NHS Foundation Trust, Warrington, UK. [233]Ege University Hospital, Department of Paediatric Hematology-Oncology, Izmir, Turkey. [234]University Hospitals Coventry and Warwickshire, Coventry, UK. [235]Infection, Immunity and Inflammation, University of Leicester, Leicester, UK. [236]School of Paediatrics and Child Health, University of Western Australia, Perth, Australia. [237]Genetic Services of Western Australia, Western Australian Register of Developmental Anomalies and Office of Population Health Genomics, Public and Aboriginal Health Division, Western Australian Department of Health, Perth, Australia. [238]Genetic and Rare Diseases Program, Telethon Kids Institute, Perth, Australia. [239]University Hospitals Plymouth NHS Trust, Plymouth, UK. [240]Gloucestershire Royal Hospital, Gloucestershire Hospitals NHS Foundation Trust, Gloucester, UK. [241]North West Thames Regional Genetic Service, London North West University Healthcare NHS Trust, Harrow, UK. [242]Department of Neurology, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [243]NHS, NHS Trust, UK. [244]Norfolk and Norwich University Hospitals NHS Foundation Trust, Norwich, UK. [245]Blizard Institute, Barts and The London School of Medicine & Dentistry, Queen Mary University of London, London, UK. [246]Royal Hallamshire Hospital NHS Foundation Trust, Sheffield, UK. [247]Royal Surrey County Hospital NHS Foundation Trust, Guildford, UK. [248]Colchester Hospital University NHS Foundation Trust, Colchester, UK. [249]Medical University of Vienna, Vienna, Austria. [250]National Ehlers–Danlos Syndrome Diagnostic Service, Northwick Park Hospital, London, UK. [251]Southampton Children's Hospital, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [252]Faculty of Medicine, University of Southampton, Southampton, UK. [253]West Middlesex University Hospital, Chelsea and Westminster Hospital NHS Foundation Trust, London, UK. [254]Freeman Hospital, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [255]University Hospital of Wales, Cardiff, UK. [256]Department of Neurology, King's College Hospital NHS Foundation Trust, London, UK. [257]Maastricht University Medical Centre, Maastricht, The Netherlands. [258]Royal Manchester Children's Hospital, Manchester University NHS Foundation Trust, Manchester, UK. [259]Lister Hospital, East and North Hertfordshire NHS Trust, Stevenage, UK. [260]Chelsea and Westminster Hospital NHS Foundation Trust, London, UK. [261]Queen Charlotte's and Chelsea Hospital, Imperial College Healthcare NHS Trust, Du Cane Road, London, UK. [262]Alder Hey Children's Hospital, Liverpool, UK. [263]Barnet General Hospital, Royal Free London NHS Foundation Trust, London, UK. [264]Leeds General Infirmary, Leeds Teaching Hospitals NHS Trust, Leeds, UK. [265]Birmingham Women's Hospital, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK. [266]Sussex Kidney Unit, Royal Sussex County Hospital, Brighton and Sussex University Hospitals, Brighton, UK. [267]Addenbrooke's Treatment Centre, Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [268]Renal Medicine, Leeds Teaching Hospitals NHS Trust, Leeds, UK. [269]Middlemore Hospital, Auckland, New Zealand. [270]School of Life Sciences, University of Lincoln, Lincoln, UK. [271]Molecular and Clinical Sciences Research Institute, St George's University of London, London, UK. [272]Nuffield Department of Women's and Reproductive Health, Oxford University Hospitals NHS Trust, Oxford, UK. [273]St George's University Hospitals NHS Foundation Trust, London, UK. [274]Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK. [275]Birmingham Children's Hospital, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK. [276]Royal Devon and Exeter NHS Foundation Trust, Exeter, UK. [277]Department of Clinical Immunology, Hopital Saint-Louis, Assistance Publique-Hopitaux de Paris, University Paris Diderot, Sorbonne Paris Cite, Paris, France. [278]UCL MRC Laboratory for Molecular Cell Biology, London, UK. [279]Barts and The London School of Medicine and Dentistry, Haemophilia Centre, The Royal London Hospital, London, UK. [280]Dept of Haematology, Sheffield Children's Hospital NHS Foundation Trust, Sheffield, UK. [281]Cardiff University, Cardiff, UK. [282]Immunodeficiency Centre for Wales, Heath Hospital, Cardiff, UK. [283]Department of Molecular Medicine, Sapienza University of Rome, Rome, Italy. [284]The Princess Alexandra Hospital NHS Trust, Harlow, UK. [285]Shaare Zedek Medical Center, affiliated with Hebrew-University Medical School, Jerusalem, Israel. [286]Leeds Children's Hospital, The Leeds Teaching Hospitals NHS Trust, Leeds, UK. [287]Department of Internal Medicine, Eccles Institute of Human Genetics, University of Utah Health Sciences Center, Salt Lake City, USA. [288]Stroke Prevention Research Unit, University of Oxford, Oxford, UK. [289]Experimental Biomedicine, University Hospital Würzburg, Würzburg, Germany. [290]Institute of Cardiovascular Research Royal Holloway University of London (ICR2UL), London, UK. [291]Epsom General Hospital, Epsom, UK. [292]Faculty of Life Sciences and Medicine, King's College London, London, UK. [293]Ninewells Hospital and Medical School, NHS Tayside, Dundee, UK. [294]Aberdeen Royal Infirmary, NHS Grampian, Aberdeen, UK. [295]Haematology Department, Manchester Royal Infirmary, Central Manchester University Hospitals National Health Service Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK. [296]Sheffield Children's Hospital NHS Foundation Trust, Sheffield, UK.